Journal of Computer Applications
计算机应用, 2020, 40(7): 1996 - 2002
ISSN 1001-9081
CODEN JYIIDU
2020-07-10
http://www.joca.cn

文章编号:1001-9081(2020)07-1996-07
DOI:10.11772/j.issn.1001-9081.2019111915

改进粗糙集属性约简结合*K*-means聚类的网络入侵检测方法

王　磊*
(苏州大学 信息化建设与管理中心,江苏 苏州 215006)
(*通信作者电子邮箱 wanglei01005@163.com)

摘　要:面对日益复杂的网络环境,传统入侵检测方法误报率高、检测效率低,且存在优化过程中准确性和可解释性相互矛盾等问题,因此提出一种结合改进粗糙集属性约简和*K*-means聚类的网络入侵检测(IRSAR-KCANID)方法。首先基于模糊粗糙集属性约简对数据集进行预处理,优化异常的入侵检测特征;再利用改进*K*-means聚类算法估计入侵范围阈值,并对网络特征进行分类;然后根据用于特征优化的线性规范相关性,从所选择的最优特征探索特征关联影响尺度以形成特征关联影响量表,完成对异常网络入侵的检测。实验结果表明,特征优化聚类后的最小化测量特征关联影响量表能在保证最大预测精度的前提下,最小化入侵检测过程的复杂度并缩短完成时间。

关键词:网络异常检测;改进粗糙集属性约简;改进*K*-means聚类;相关性分析;特征关联尺度
中图分类号:TP391　　**文献标志码**:A

Network intrusion detection method based on improved rough set attribute reduction and *K*-means clustering

WANG Lei*
(*Center of Information Development and Management*, *Soochow University*, *Suzhou Jiangsu* 215006, *China*)

Abstract: Under increasingly complex network environment, traditional intrusion detection methods have high false alarm rate, low detection efficiency and the contradiction between accuracy and interpretability in the optimization process. Therefore, an Improved Rough Set Attribute Reduction and optimized *K*-means Clustering Approach for Network Intrusion Detection (IRSAR-KCANID) was proposed. Firstly, the dataset was preprocessed based on the attribute reduction of fuzzy rough set in order to optimize the anomalous intrusion detection features. Then, the threshold of intrusion range was estimated by improved *K*-means clustering algorithm, and the network features were classified. After that, according to the linear canonical correlation used for feature optimization, the feature association impact scale was explored from the selected optimal features in order to form the table of feature association impact scale, and the detection of anomalous network intrusion was completed. The experimental results show that the minimum measured feature association impact scale table after feature optimization clustering can minimize the complexity of intrusion detection process and shorten the completion time on the premise of guaranteeing maximum prediction accuracy.

Key words: network anomaly detection; improved rough set attribute reduction; improved *K*-means clustering; correlation analysis; feature association scale

0　引言

网络安全问题一直是全社会关注的焦点,随着网络环境的日益复杂,包括防火墙、安全路由及数据加密等静态网络安全保护方法已很难满足人们对于网络安全性能的需求。

入侵检测系统(Intrusion Detection System, IDS)作为一种网络安全主动防御技术,能够对防火墙等传统安全保护体系起到辅助作用[1],通过监控流经某个节点的流量,实现对入侵行为的检测,并生成报警信号发送至系统管理员,典型的IDS通常包括事件采集、事件分析和事件响应三个核心环节,其检测方法主要可分为两种类型:误用IDS和异常IDS。现有IDS均或多或少存在有效性低、适应性不强、误报率高以及可扩展性不高等问题。其中:误用IDS根据已知攻击和系统弱点的参数识别入侵,然而它无法识别新的或不熟悉的攻击类型;异常IDS则基于正常行为的参数,并使用它们来识别任何与正常行为相差甚远的行为[2]。误用入侵检测的机制是训练现有的入侵模式,并将考虑用于检查的数据,与先前的模式相匹配,以识别入侵。IDS一般挂接在所有所关注流量都必须流经的链路上,而所关注流量则是指来自高危网络区域的访问数据和需要进行统计、监视的网络报文数据。即无论是误用IDS还是异常IDS,都离不开对数据的挖掘与处理。

利用数据挖掘技术开发的IDS通常具有检测网络入侵的优异性能和泛化能力,从而使其具有高效的入侵检测性能。然而,实现和安装这种系统的过程是复杂的,系统的固有复杂性可以根据准确性、能力和可用性的参数,组织成单独的问题集[3]。与使用数据挖掘技术构建的IDS相关联的一个关键问题主要是基于异常检测的那些技术,与先前基于手工签名的检测技术相比,其误报率更高[4]。因此,对于这些技术来说,

收稿日期:2019-11-11;**修回日期**:2020-04-10;**录用日期**:2020-04-12。　**基金项目**:国家自然科学基金青年科学基金资助项目(61802272)。
作者简介:王磊(1987—),男,江苏苏州人,研究员,硕士,主要研究方向:云计算、计算机网络安全防护。

审计数据的处理和在线入侵的检测比较困难,并且需要大量的训练数据。文献[5]提出了一种结合了统计技术和自组织映射来检测网络中异常的分类方法(Statistical Techniques and Self-organizing Maps, STSM),其中主成分分析(Principal Component Analysis, PCA)和Fisher判别比用于特征选择和噪声消除,概率自组织映射用于将网络事务分类为正常或异常。文献[6]提出了一种结合数据挖掘方法的混合技术(Hybrid Technique that combines Data Mining Approaches, HT-DMA)。该方法中,K-means聚类算法用于减少与每个数据点相关联属性的数量,再将支持向量机(Support Vector Machine, SVM)的径向基函数(Radial Basis Function, RBF)用于异常网络入侵检测。文献[7]提出了基于距离和的SVM混合学习(Distance Sum-based SVM, DSSVM)方法,用于建模有效的IDS。在DSSVM中,获得基于每个数据样本与数据集中的聚类中心特征维度之间的相关性的距离和,并将SVM用作分类器。

然而现有方法需要大量的训练数据,并且与系统的学习过程相关的复杂性很高。因此提出一种基于改进粗糙集属性约简和K-means聚类的网络入侵检测方法(Improved Rough Set Attribute Reduction and optimized K-means Clustering Approach for Network Intrusion Detection, IRSAR-KCANID)。所提方法首先基于改进模糊粗糙集属性约简对数据集进行预处理,优化异常的入侵检测特征,然后利用改进K-means聚类算法进行入侵检测特征分析和入侵范围估计阈值估计,并对网络特征进行分类;再根据用于特征优化的线性规范相关性,从所选择的最优特征探索关联影响尺度,形成特征关联影响量(Feature Association Impact Scale, FAIS)表,完成对异常网络入侵的快速准确检测。主要创新体现在以下几个方面:

1)现有方法在入侵检测数据训练方面耗时较多,提出的方法利用改进模糊粗糙集属性约简对数据集进行了预处理,优化异常的入侵检测特征,避免了对大量数据的训练,缩短了入侵检测时间;

2)现有大多数入侵检测方法仅仅是发现攻击行为,没有对攻击进行有效的分类,提出的方法在数据预处理的基础上,利用改进K-means聚类算法进行入侵检测特征分析和入侵范围估计阈值估计,并对网络特征进行分类。

3)在聚类结果的基础上,根据用于特征优化的线性规范相关性,从所选择的最优特征探索关联影响尺度形成关联影响量表,从而完成对异常网络入侵的检测。

特征相关性实验结果表明,特征优化聚类后的最小化测量特征关联影响量表能在保证最大预测精度的前提下,最小化入侵检测过程的复杂度并缩短完成时间。

1　基于改进粗糙集属性约简的数据集预处理

由于原始数据往往包含隐含信息[8-9],本文利用改进粗糙集属性约简(Improved Rough Set Attribute Reduction, IRSAR)将这些隐含信息提取出来,在保留原始特征的同时更好地表现数据特征。将网络连接记录表示为四元组$FS=(U,A_t,V,f)$,其中:U为整个网络数据集;A_t是一个非空的有限属性集,t表示属性集数量;$V=\bigcup_{a\in A_t}V_a$表示属性a域集合;$f=U\times A_t$表示信息函数。

由于传统的粗糙集理论只能处理离散属性集,无法很好地处理包含大量连续值的网络连接数据[10-11],因此引入模糊理论,利用模糊粗糙集的信息增益率对网络连接数据特征进行自动选取。

将引入模糊理论的网络连接记录表示为$FIS=(U,C\cup D,V,f)$,设$B\subseteq C,\forall a\in C-B$,$C$为条件属性集,$B$为约简的属性集,$D$为决策属性集,属性$a$的信息增益率为:

$$Gain_{\mathrm{Ratlo}}(a,B,D)=\frac{I(B\cup\{a\};\ D)-I(B;\ D)}{H(a)}\tag{1}$$

其中,$Gain_{\mathrm{Ratlo}}$表示增益率,$Gain_{\mathrm{Ratlo}}(a,B,D)$可用于衡量属性$a$的重要程度,可以通过每次选择增益率最大的特征进行属性选取,最终获得的属性集即为约简的本征属性集。IRSAR的数据集预处理主要步骤如下,其中输入为数据集X、条件属性集C、决策属性集D,输出为约简的属性集B:

1)清空B集合,计算$Gain_{\mathrm{Ratlo}}(a,B,D)$,并筛选其最大值;

2)如果$\max Gain_{\mathrm{Ratlo}}(a,B,D)>0$,则$B\leftarrow B\cup\{a\}$,返回1);

3)集合B为属性约简后的属性集合。

模糊等价关系是模糊粗糙集的核心,假如给定非空有限数据集X,X上的模糊等价关系R可以用关系矩阵$\boldsymbol{M}_r$表示为:

$$\boldsymbol{M}_r=\begin{bmatrix}r_{11}&r_{12}&\cdots&r_{1n}\\r_{21}&r_{22}&\cdots&r_{2n}\\\vdots&\vdots&&\vdots\\r_{n1}&r_{n2}&\cdots&r_{nn}\end{bmatrix}\tag{2}$$

其中$r_{ij}\in[0,1]$是x_i与x_j的关联值;x_i和x_j分别表示不同数据在同一属性上的值,$x_i,x_j\in X$,模糊等价关系需要满足自反、对称和传递性,能够实现信息增益率对网络连接数据特征属性集进行自动筛选,以获得约简的本征属性集,从而有效提高入侵检测算法的稳定性。相较于经典粗糙集理论只能处理离散属性集的短板,改进粗糙集属性能够获得保留原始特征辨别能力的属性子集,能够很好地处理包含大量连续值的网络连接数据。

2　特征分析与影响尺度阈值估计方法

2.1　K-means聚类及其改进

K-means聚类算法采用评价指标来度量距离的相似性[12-13],其主要思想体现为以下三点:

1)在样本数据中,样本数量为k,且为任意设设定,设定的样本代表一个簇的初始中心或者均值;

2)数据样本与每个聚类中心之间的距离通常用欧氏距离公式计算,每个数据样本根据计算结果被分配到最近的类;

3)调整聚类中心并对得到的新类进行再次计算,聚类准则函数收敛的条件是聚类中心不再变化,即可终止对样本数据的聚类调整,从而结束算法。

改进K-means算法则针对初值选取敏感问题,算法中簇心的初始位置在算法开始时通过临时指定,再通过样本数据各维度的最大值和最小值计算,结合多次迭代来选取最佳的簇心,期间采用随机梯度下降的方法来取代批量梯度下降以防止K-means算法陷入局部最优。假定$h(\theta)$为所需要拟合的函数,$J(\theta)$为损失函数,其函数形式分别表示为:

$$h(\theta)=\sum_{j=0}^{m}\theta_jX_j\tag{3}$$

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(Y^i-h_\theta(Y^t))^2\tag{4}$$

其中：m 表示训练集的数量，θ 表示多次迭代计算所需要求取的值，X 和 Y 为数据集，i 表示迭代计数，t 为损失因子，参数个数表示为 j。当求解出 θ 时最终要拟合的函数 $h(\theta)$ 的值也相应求得。

损失函数也可以改写为：

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(y^i-h_\theta(x^t))^2=\frac{1}{m}\sum_{i=1}^{m}cost(\theta,(x^t,y^i)) \tag{5}$$

其中 $cost(\theta,(x^t,y^i))$ 可表示为：

$$cost(\theta,(x^t,y^i))=\frac{1}{2}(y^i-h_\theta(x^i))^2 \tag{6}$$

此处损失函数所对应的辨识训练集中每个样本数据的隶属度，对于每个样本数据的损失函数，通过对 θ 求偏导可以求出相应的梯度，其中 θ 可以根据以下公式更新：

$$\theta_j'=\theta_j+(y^i-h_\theta(x^i))x_j^i \tag{7}$$

在计算过程中 θ 可以通过迭代计算不断更新，但如果学习效率设置过高则可能导致振荡现象。因此可以引进学习率 α 进行改进，若假设 $f(\alpha)=h(\boldsymbol{x}_k+\alpha\boldsymbol{d}_k)$，其中当前样本点设置为 $\boldsymbol{x}_k$，搜索方向设置为 $\boldsymbol{d}_k$，则可得随机梯度下降过程所寻找的 $f(\alpha)$ 最小值为：

$$\alpha=\arg\min_{\alpha>0}f(\alpha)=\arg\min_{\alpha>0}(\boldsymbol{x}_k+\alpha\boldsymbol{d}_k) \tag{8}$$

对学习率的函数导数的分析：若 α=0，则有

$$f'(0)=\nabla h(\boldsymbol{x}_k+0*\boldsymbol{d}_k)^{\mathrm{T}}\boldsymbol{d}_k=\nabla h(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{d}_k \tag{9}$$

下降方向 $\boldsymbol{d}_k$ 可以选负梯度方向 $\boldsymbol{d}_k=-\nabla h(\boldsymbol{x}_k)$，从而使 $f'(0)>0$。假如找到的 α 足够大，并且使得 $f'(\hat{\alpha})>0$，则一定存在某个 α，使得 $f'(\alpha^*)>0$，其中 α^* 即为改进设置的学习率。

改进 *K*-means 聚类算法工作步骤如下，输入 k（簇数），输出标记好的 k 个簇集合。

1）手动设定 k 个临时簇心；

2）在样本数据每个向量的维度以及各自维度最大值和最小值选取簇心；

3）根据选取的样本数据 X_i 找出距离它最近的簇心，并把簇心向 X_i 方向移动；

4）每次移动数据项时都乘以学习率 α，其变化趋势随迭代次数增加而不断减小；

5）返回步骤2）；

6）对簇心进行更新；

7）直到簇心位置固定不变；

8）根据数量以及标记判别该簇正常与否。

改进后的 *K*-means 算法对于初值选取要求有所降低，相较于原始算法簇心的初始位置可以在算法开始时临时指定，无需进行繁琐的初值整定；此外，改进算法在稳定性方面也有一定的提升，因为学习率 α 的设置改进，可以避免因学习效率设置过高而导致的振荡现象。

2.2 入侵检测特征分析与特征关联影响尺度阈值估计

2.2.1 入侵检测特征分析

网络事务集包含的42个特征可以分为连续和分类的值，为了便于优化，需要将所有最初字母及连续数值转换为分类。预处理的一组网络事务根据其标签进行分区，使得正常事务是一组，拒绝服务（Denial of Service，DoS）攻击事务是另一组。

将字母数字值表示为数值，并将联系续值表示为分类值，其具体步骤如下：

1）考虑具有字母数字值的每个要素，然后列出所有可能的唯一值，并使用从1开始的增量索引列出它们；

2）用适当的索引替换值；

3）考虑具有连续值的每个要素，然后将它们划分为一组具有最小值和最大值的范围，以便事件在所有这些范围内均匀分布。

考虑结果正常交易集（Normal Trade Set，*NTS*）中的每个特征值集合 $f_iv(NTS)$ 及其覆盖百分比为 $f_iv=\{f_i(v_1,c_1),f_i(v_2,c_2),\cdots,f_i(v_j,c_j)\}$，$v$，$c$ 为特征量，然后，可以按照以下步骤中的描述执行每个攻击 A 的特征优化：

1）考虑交易集 $ts(A_k)$ 表示攻击类型 A_k（假设为DoS攻击）。

2）对于每个特征 $f_i(A_k)$，将所有值视为集合 $f_iv(A_k)$。创建大小为 $\left|f_iv(A_k)\right|$ 的空集 $\overline{f_iv}$，并根据其覆盖百分比填充 f_iv 中的值，使得 $\left|f_iv(A_k)\right|\cong\left|\left|f_iv\right|\right|$。这里 $\left|f_iv(A_k)\right|$ 表示 $f_i(A_k)$ 的特征值集的大小。

3）求解网络测试系统的特征值 $\overline{f_iv}$，使得 $\overline{f_iv}$ 与 $f_iv(A_k)$ 大小兼容，并且还能表示 $f_iv(NTS)$ 中的值的覆盖率。

4）此过程应适用于攻击 A_k 的网络事务中设置的所有特征值。

5）找出 $f_iv(A_k)$ 和 $\overline{f_iv}$ 之间的典型相关性。如果得到的典型相关性小于给定阈值或零，那么特征 $f_i(A_k)$ 可以被认为是评估入侵范围规模的最佳值。

根据上述步骤中说明的过程，可以识别特定攻击 A_k 的最佳特征。

2.2.2 特征关联影响尺度阈值估计

通过聚合 $\boldsymbol{A}$ 的每一行来找到特权权重（将形成表示特权权重 $\boldsymbol{v}$），再通过 $\boldsymbol{A}$ 和 $\boldsymbol{v}$ 之间的乘法找到枢轴权重：

$$\boldsymbol{u}=\boldsymbol{A}\times\boldsymbol{v} \tag{10}$$

那么特征分类值 f_iv_j 的尺度阈值 *fas* 可以通过如下公式计算：

$$fas(f_iv_j)=\frac{\sum_{k=1}^{|STVS|}\{u(tvs_k):(f_iv_j\rightarrow tvs_k)\neq 0\}}{\sum_{k=1}^{|STVS|}u(tvs_k)} \tag{11}$$

特征分类值 f_iv_j 和 $f_{i'}v_{j'}$ 之间的 *fas* 可以表示为：

$$fas(f_iv_j\leftrightarrow f_{i'}v_{j'})=\frac{\sum_{k=1}^{|STVS|}\{u(tvs_k)\exists(f_iv_j,f_{i'}v_{j'})\subset tvs_k\}}{\sum_{k=1}^{|STVS|}u(tvs_k)} \tag{12}$$

其中：tvs_k 表示 k 交易价值集，$|STVS|$ 表示事务值集的总数。

另外，每个交易价值集 tvs_i 的特征关联影响量表 *fais* 和 *faist* 阈值可以分别表示为：

$$fais(tvs_i)=1-\frac{\sum_{j=1}^{|m|}\{fas(val_j)\exists val_j\in V\}):(val_j\subset tvs_i)\}}{\left|tvs_i\right|} \tag{13}$$

$$faist=\sum_{i=1}^{|STVS|}fais(tvs_i)\Big/|STVS| \tag{14}$$

其中：$val_j\in V$ 表示特征差值。

每个交易价值 *faist* 的标准差需要进一步测量集合，以估计 *faist* 阈值的上下限和挑战黑洞（Challenge Collapsar，CC）阈

值范围。其中，cc阈值是*faist*的一个临界值；下限为cc平均值与cc标准差之间的差值，上限为cc平均值与cc标准差之和。阈值设定的目的在于对以上三种范围进行阈值额定，与此对应的范围分别为不相关性、弱相似性和强相似性。发现的正常记录总数为测试数据记录的总和，估算标准偏差表示如下：

$$sdv_{faist}=\sqrt{\left(\sum_{i=1}^{|STVS|}fais(tvs_i)-faist^2\right)\Big/(|STVS|-1)} \tag{15}$$

*faist*系列可以探索范围如下：

*faist*范围的下限是：

$$faist_1=faist-sdv_{faist} \tag{16}$$

*faist*范围的上限是：

$$faist_h=faist+sdv_{faist} \tag{17}$$

当且仅当$fais(nt)<faist_1$时，网络事务*nt*可以说是安全的。

通过对网络中不同标注下数据进行处理，结合模糊等价关系矩阵，可获得输入信号参数入侵特征阈值的参考指标集如下：

$$\boldsymbol{M}_g=\begin{bmatrix} r'_{11} & r'_{12} & \cdots & r'_{1n} \\ r'_{21} & r'_{22} & \cdots & r'_{2n} \\ \vdots & \vdots & & \vdots \\ r'_{n1} & r'_{n2} & \cdots & r'_{nn} \end{bmatrix} \tag{18}$$

通过上式构建$\boldsymbol{M}_g$关联模型，并通过不断训练改变参数个数与入侵特征阈值，获取异常度量关联矩阵：

$$\boldsymbol{M}_m=\sum m\boldsymbol{M}_g \tag{19}$$

其中m表示参数个数，则有入侵检测特征关联影响阈值为：

$$T_g=Ave(faist-\boldsymbol{M}_m) \tag{20}$$

2.3　数据集特征相关性分析并聚类

考虑两个多维数据集X和Y，并且利用基于标准统计技术的典型相关分析（Canonical Correlation Analysis，CCA），利用二阶的自协方差和互协方差矩阵，建立数据集之间的线性关系。该技术基于两个基础，每个基础用于数据集X和Y，其中互相关矩阵变为对角线，并且对角线的相关性最大化。

研究用于实现规范相关的参数，其中，X和Y应该相等；然而，假设平均值为零，数据向量$\boldsymbol{x}\in X$和$\boldsymbol{y}\in Y$可以具有变化的尺寸。使用特征向量方程求解规范相关计算：

$$\begin{cases}\boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{w}_x=r^2\boldsymbol{w}_x \\ \boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xy}\boldsymbol{w}_y=r^2\boldsymbol{w}_y\end{cases} \tag{21}$$

这里，$\boldsymbol{C}_{xx}$、$\boldsymbol{C}_{xy}$、$\boldsymbol{C}_{yy}$、$\boldsymbol{C}_{yx}$均为交叉协方差矩阵，其中r^2本征值是规范相关的平方，$\boldsymbol{w}_x$和$\boldsymbol{w}_y$是归一化CCA基矢量。方程的解等价于非零值，其数量等于$\boldsymbol{x}$和$\boldsymbol{y}$，表示考虑具有较小维数值的数据向量。当$\boldsymbol{C}_{yx}=\boldsymbol{C}_{xy}^{\mathrm{T}}$时，式(21)被转换为：

$$\begin{cases}\boldsymbol{C}_{xy}\boldsymbol{C}_{xy}^{\mathrm{T}}\boldsymbol{w}_x=r^2\boldsymbol{w}_x \\ \boldsymbol{C}_{yx}\boldsymbol{C}_{yx}^{\mathrm{T}}\boldsymbol{w}_y=r^2\boldsymbol{w}_y\end{cases} \tag{22}$$

这些方程描述了交叉协方差矩阵$\boldsymbol{C}_{xy}$的奇异值分解：

$$\boldsymbol{C}_{xy}=\boldsymbol{U}\sum\boldsymbol{V}^{\mathrm{T}}=\sum_{i=1}^{L}r_i\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}} \tag{23}$$

这里$\boldsymbol{U}$和$\boldsymbol{V}$表示包括奇异向量$\boldsymbol{u}_i$和$\boldsymbol{v}_i$的正交平方矩阵。$\boldsymbol{w}_x$和$\boldsymbol{w}_y$表示传递规范相关性的基础向量。矩阵$\boldsymbol{U}$和$\boldsymbol{V}$以及$\boldsymbol{u}_i$和$\boldsymbol{v}_i$的向量维度通常根据$\boldsymbol{x}$和$\boldsymbol{y}$数据向量的维度变化而变化。

$$\boldsymbol{Q}=\begin{bmatrix}\boldsymbol{D} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} \tag{24}$$

伪对角矩阵$\boldsymbol{Q}$由对角矩阵$\boldsymbol{D}$和附加零矩阵构建，这将使得矩阵$\boldsymbol{Q}$与$\boldsymbol{x}$，$\boldsymbol{y}$各维度兼容。如果$\boldsymbol{C}_{xy}$具有满秩，则非零奇异值基本上是非零规范相关，其数量小于$\boldsymbol{x}$和$\boldsymbol{y}$数据矢量维度中的任何一个。

3　特征关联影响量表的入侵检测

测量特征关联支持度量的方法是将给定训练集的网络事务记录和在这些网络事务中使用的特征分类值视为两个独立集合，并进一步构建这两者之间的双工图[14]。所提入侵检测基于以下理想性假设和操作步骤实施。

3.1　理想性假设

特征$\{f_1,f_2,\cdots,f_n\forall f_i=\{f_iv_1,f_iv_2,\cdots,f_iv_m\}\}$是对特定攻击$A_k$是最佳的分类值，通过应用于网络事务集$T(A_k)$的典型相关分析来选择。这里$T(A_k)$是给定训练集的特定攻击$A_k$的网络事务记录集，使得：$T=\{t_1,t_2,\cdots,t_n\forall t_i=\{val(f_1),val(f_2),\cdots,val(f_i),val(f_{i+1}),\cdots,val(f_n)\}\}$属于每个网络事务特征的分类值集合，称为事务值集合*tvs*，并且将所有事务值集合称为*STVS*。在上面的描述中，$val(f_i)$可以被定义为$val(f_i)\in\{f_iv_1,f_iv_2,\cdots,f_iv_m\}$，此后，术语特征指的是特征的当前分类值。当且仅当$(val(f_i),val(f_j))\in tvs_k$时，对于两个特征$val(f_i)$和$val(f_j)$，$val(f_i)$与$val(f_j)$连接。

3.2　方法与步骤

本文通过示例探索该过程，将*STVS*要素的发散向量表示为$\boldsymbol{V}=\{val_1,val_2,\cdots,val_8\}$。在表1和图2中，每个元素$\{val_1,val_2,\cdots,val_8\}$可以是$f_iv_j$，使得$\{f_iv_j\exists i\in[1,2,\cdots,n]\wedge j\in[1,2,\cdots,m]\}$。

在检测val_k的每个特征分类值f_iv_j与网络事务记录的关联过程中，需要在*STVS*和特征分类值之间建立双工图。

形成双重图可认为图关系是二分的，并且在特征和事务值集之间形成边。此图中的每个关系都表示特征对网络事务的作用[15]。当且仅当该特征f是*tvs*的一部分时，交易值集合*tvs*和特征f之间的边缘才存在可能，这可以表示为$e_{tvs\leftarrow f}\exists f\in tvs$。

表1　*STVS*和特征分类值之间关联的二进制表示

Tab. 1　Binary representation of correlation between *STVS* and feature classification value

STVS	val_1	val_2	val_3	val_4	val_5	val_6	val_7	val_8	结果表示
tvs_1	0	1	0	0	1	1	0	1	(val_2，val_5，val_6，val_8)
tvs_2	1	0	0	0		1	0	1	(val_1，val_6，val_8)
tvs_3	1	0	1	0	0	0	1	0	(val_1，val_3，val_7)
tvs_4	0	1	0	0	0	0	1	0	(val_2，val_7)
tvs_5	1	0	0	1	0	1	1	1	(val_1，val_4，val_6，val_7，val_8)
tvs_6	1	1	1	1	0	0	0	1	(val_1，val_2，val_3，val_4，val_8)

图 1 所示为加权无向图，其中特征值作为特征值之间的顶点和边。

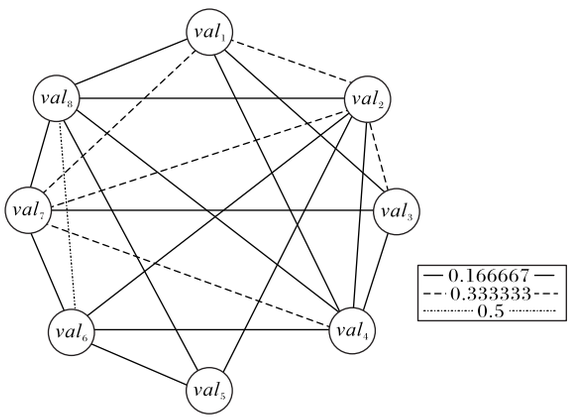

图 1 计数为 8 的分类值集示例加权图

Fig. 1 Weighted graph example of classification value set with counting of 8

任意两个特征 $val(f_1), val(f_2)$ 之间的边将按如下方式加权：

$ctvs=0$

foreach $\{tvs \forall tvs \in STVS\}$

$$ctvs+=\{1 \forall (val(f_1), val(f_2)) \subseteq tvs\} \tag{25}$$

在上面的等式中，$ctvs$ 表示事务计数，其中包含两个特征 $val(f_1)$、$val(f_2)$。然后特征 $val(f_1)$、$val(f_2)$ 之间的边缘重量可以如下测量：

$$w(val(f_1) \leftrightarrow val(f_2)) = ctvs / |STVS| \tag{26}$$

在构建加权图的过程中，本文认为当且仅当 $ctvs \geqslant 1$ 时，任何两个特征之间存在边际。

在如图 2 所示的双工图中，虚线表示连接元素属于双工图的相同级别，实线表示特征值和事务值集之间的关系。

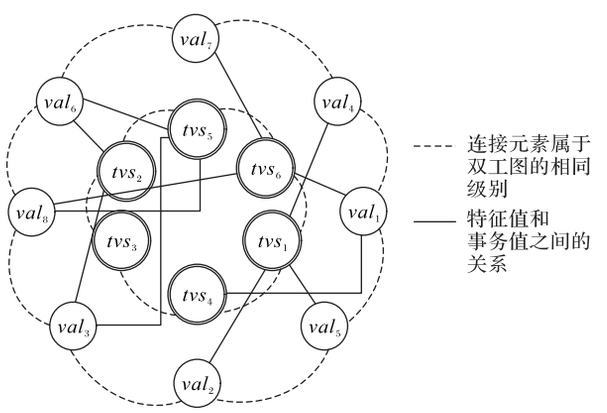

图 2 $STVS$ 和 $\boldsymbol{V}$ 之间的双工图

Fig. 2 Duplex diagram between $STVS$ and $\boldsymbol{V}$

如果在 tvs_1 中存在称为 val_1 的特征分类值 $f_i v_j$，则 val_1 和 tvs_1 之间的连接的权重将是 val_1 与在加权中定义的 tvs_1 的每个特征分类值 $\{f_i v_j \exists f_i v_j \in tvs_1\}$ 之间边的权重的总和图形[16]。

此外，将形成矩阵 $\boldsymbol{A}$，表示交易值集和特征分类值之间的双重图的边缘权重。然后获得 $\boldsymbol{A}'$，表示矩阵 $\boldsymbol{A}$ 的转置[17]。

将 $STVS$ 视为数据库，并将其描述为双工图而不会丢失信息。设 $STVS = \{tvs_1, tvs_2, \cdots, tvs_6\}$ 是事务值集的列表，$\boldsymbol{V} = \{val_1, val_2, \cdots, val_8\}$ 是相应的特征集分类值。那么，显然 $STVS$ 相当于双工图 $DG = (STVS, V, E)$。其中，特征值分类值能够跟随通道业务变化而动态调整，从而达到辨识策略的修正，实现通信网络入侵的在线监测。

这里，$E = \{tvs_i, val_i): val_i \in tvs_i, tvs_i \in STVS,\ val_i \in \boldsymbol{V}\}$。

假设给定双工图的交易值集，作为枢轴并且特征分类值作为纯特权，则可以测量枢轴和特权值[18-19]。如果在交易值集合中存在特征分类值 val_1，那么 val_1 和 tvs_1 之间的连接的权重，将是 val_1 与电视的每个特征分类值 $\{val_i \exists val_i \in tvs_1\}$ 之间的边缘权重的总和。这些权重是边缘权重，用加权图（Weighted Graph，WG）表示。根据 2.2 节所述入侵范围估计方法，对特征关联影响尺度阈值进行估计。

所提方法首先对数据集进行预处理，优化异常的入侵检测特征，然后利用改进 K-means 聚类算法估计入侵范围阈值并对网络特征进行最终分类；再根据用于特征优化的线性规范相关性，从所选择的最优特征探索特征关联影响尺度，形成特征关联影响量表，完成对异常网络入侵的检测。其具体流程如图 3 所示。

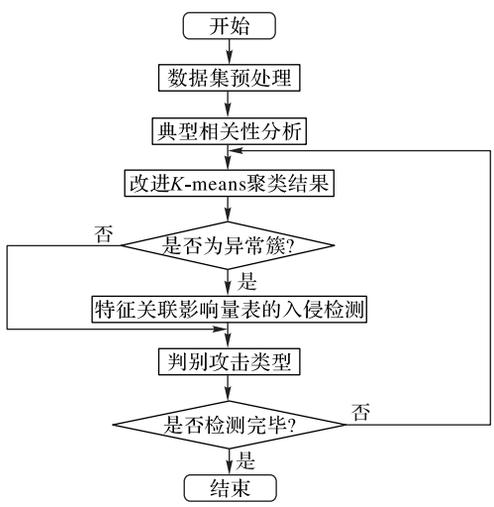

图 3 基于改进 K-means 结合关联影响尺度分析的入侵检测方法流程

Fig. 3 Flowchart of intrusion detection method based on improved K-means and association impact scale analysis

4 实验结果与分析

入侵检测评估程序生成的数据用于构建原始 KDD-99 数据集，包含接近 4 900 000 个唯一连接向量，其中每个连接向量由 41 个特征组成，34 个是连续特征，7 个是离散的特征。此外，本文还利用 CICIDS2017 通用数据集进行了对比实验，CICIDS2017 数据集是加拿大网络安全研究所于 2017 年开源的入侵检测和入侵预防数据集，通过攻击本地网络来收集流量数据，在一段时间内收集正常流量和常见的攻击流量，设计真实攻击场景，具有一定的通用性和应用性。在本文的实验中模拟的攻击属于下面描述的四种类型中的任何一种。

1）DoS。DoS 攻击是一种攻击类型，攻击者通过消耗计算机或内存资源来阻止对有效用户的访问，从而使系统无法处理有效请求。DoS 攻击的例子很多，如：teardrop、neptune、ping of death（pod）、mail bomb、back、smurf 和 land。

2）用户到根式攻击（Users-to-Root attack，U2R）。根攻击

是一种攻击类型,攻击者可以访问系统中的有效用户账户,并根据现有的系统弱点获取对系统根组件的访问权限。有几种类型的U2R攻击,例如:负载模块、缓冲区溢出、rootkit、purl。

3)远程到本地攻击(Remote-to-Local attack, R2L)。远程到本地攻击是一种攻击,其中没有账户的攻击者根据现有的计算机漏洞在本地访问合法用户账户。R2L攻击类型有:phf、warezmaster、warezclient、spy、imap、ftp_write、multihop 和 guess_passwd。

4) 探测攻击(Probing attack, PROBE)。探测攻击是一种攻击类型,攻击者会避开安防系统收集网络中计算机上的数据。PROBE 攻击类型有:nmap、satan、ipsweep 和 portsweep。在NSL-KDD数据集中,考虑的协议是TCP、UDP和ICMP。

本实验基于 Intel Core i5-5430M CPU @ 2.70 GB,4 GB RAM计算机平台,并在Linux系统中采用C程序对数据集进行预处理操作,同时采用Java执行数据分类和入侵检测,采用粗糙集工具RSES(Rough Set Exploration System)。实验通过与文献[5]和文献[7]所提方法(即STSM和DSSVM)进行对比,从入侵检测精度以及检测完成时间等方面比较了所提入侵检测方法的可行性和先进性。同时在原始KDD-99数据集实验基础上,增加了CICIDS2017通用数据集的对照实验,以验证所提方法的普适性。其中,假设网络中发生的真实的攻击事件数量M, IDS漏报的事件数量为N,在基于原始KDD-99数据集的实验中,通过数据预处理得到的训练数据为54 675条,测试记录24 533条;基于CICIDS2017通用数据集的实验中,通过数据预处理得到的训练数据为53 687条,测试记录23 645条,实验数据分布类型和结果通过多次处理和测试得到。衡量系统性能最为重要的因素有检测率(True Positive, TP)、误报率(False Positive, FP)和漏报率(False Negative, FN)。异常网络入侵检测精度($Precision$)是入侵检测方法的主要度量指标,分析得出了入侵检测的精确度度量方法:

$$Precision = TP/(TP + FP) \tag{27}$$

其中:TP为正确识别为入侵事件与所有入侵的事件数的比值,FP为错误识别为入侵事件与所有非入侵的事件数的比值,FN为存在漏报的事件数与所有非入侵的事件数的比值。

实验将提出的方法与STSM和DSSVM在KDD-99数据集上进行了对比,其结果如图4所示。

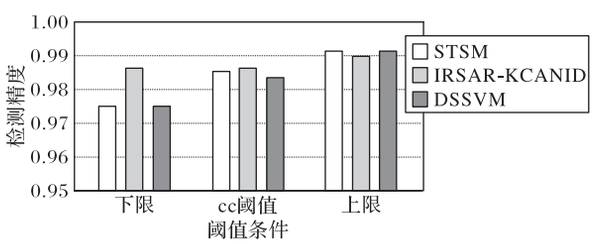

图4　KDD-99数据集上典型发散相关阈值下IRSAR-KCANID预测精度的性能分析

Fig. 4　Performance analysis of IRSAR-KCANID prediction accuracy under typical divergence correlation threshold on KDD-99 dataset

从图4中可以看出,提出的方法在阈值下限和临界阈值附近对异常网络入侵的检测精度优于STSM和DSSVM方法,其检测精度均在97%以上,但在阈值上限处的精度则比另外两种方法稍差。

同时,在同样的实验条件下,将所提方法与STSM和DSSVM在CICIDS2017数据集上也进行对比,三者的阈值设定为各自在训练集重构误差的均值。

由图5可知,在阈值下限附近所提方法对入侵检测精度明显优于STSM和DSSVM方法,且在临界阈值条件下也保持了较好的精度优势,在阈值上限条件下,三种方法大体相同,均在99%以上。

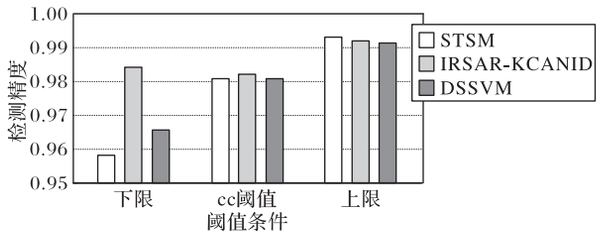

图5　CICIDS2017数据集上典型发散相关阈值下IRSAR-KCANID预测精度的性能分析

Fig. 5　Performance analysis of IRSAR-KCANID prediction accuracy under typical divergence correlation threshold on CICIDS2017 dataset

在不同标记下的不同场景典型相关性实验中,对时间复杂度进行了实验分析,提出的方法实验结果如图6所示。

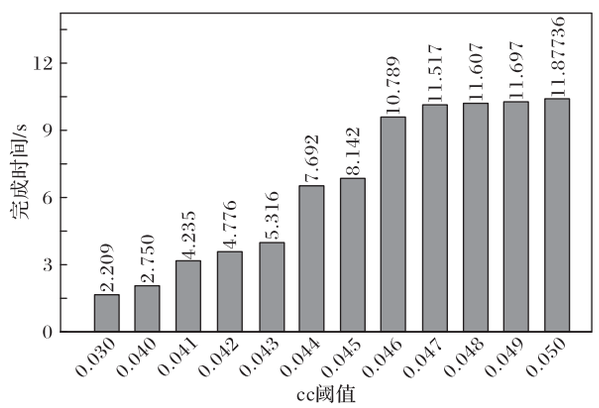

图6　在不同的典型相关阈值下IRSAR-KCANID的入侵检测完成时间

Fig. 6　Intrusion detection completion time of IRSAR-KCANID under different typical correlation thresholds

由图6可知,由于cc阈值存在变化,所需要的时间复杂度也是可缩放的。当cc阈值较小时,所需要的完成时间较少,如cc阈值为0.03时,仅需2.209 s便可完成入侵检测;随着cc阈值逐渐增大,所需要的完成时间逐渐延长,当cc阈值接近0.047时,完成时间趋于稳定时间11.6 s左右。

此外,实验将所提方法与STSM与DSSVM在不同数据集中的不同属性数量下入侵检测时间复杂度方面的对比,其实验结果如表2所示。

如表2所示,在不同数据集的同一属性数量水平下,不同数据集对入侵检测完成时间几乎没有影响。以KDD-99为例,STSM与DSSVM方法比所提的IRSAR-KCANID方法入侵检测时间更长。当属性数量为90时,STSM与DSSVM方法时间分别为0.115 s和0.095 s,而提出的方法仅为0.06 s;当属性数量为250时,STSM与DSSVM方法时间分别为0.945 s和0.935 s,提出的方法为0.324 s,大约节省60%的网络入侵检测时间;在CICIDS2017数据集中,当属性数量为70时,STSM方法时间为0.077 s,DSSVM与所提方法的时间为0.033 s;当属性数量为230时,STSM与DSSVM方法时间分别为0.943 s和0.893 s,而所提方法所需时间仅为0.535 s,相比于较快的DSSVM方法能节省大约0.0363 s入侵检测时间。由此可见,在不同的数据集中,入侵检测方法在属性数量越大时,所需要的入侵检测事例越多,所提方法相对于其他方法在不同数据

集中对于入侵检测所节约的时间成本越明显。

表2 不同属性数量下入侵检测完成时间对比 单位:s

Tab. 2 Comparison of intrusion detection completion time complexity with different attribute numbers unit:s

属性数量	STSM		DSSVM		IRSAR-KCANID	
	KDD-99	CICIDS 2017	KDD-99	CICIDS 2017	KDD-99	CICIDS2017
70	0.075	0.077	0.035	0.033	0.032	0.033
90	0.115	0.125	0.095	0.094	0.063	0.060
110	0.223	0.222	0.124	0.125	0.201	0.211
130	0.312	0.322	0.272	0.271	0.302	0.312
150	0.496	0.486	0.398	0.390	0.365	0.364
170	0.589	0.599	0.588	0.590	0.387	0.386
190	0.799	0.789	0.779	0.778	0.421	0.423
210	0.824	0.833	0.764	0.774	0.432	0.435
230	0.942	0.943	0.892	0.893	0.535	0.535
250	0.945	0.943	0.935	0.933	0.324	0.329

5 结语

本文提出的IRSAR-KCANID简化了特征分析过程,使用基准数据集进行实验,同时引入IRSAR对数据集进行预处理,采用改进*K*-means聚类方法对数据特征进行聚类分析。实验结果表明,规范相关分析对于选择用于训练的网络事务的最优属性十分重要,提出的方法在特征相关聚类的基础上,结合关联影响尺度进行入侵检测,在保证最大化检测精度的前提下,最小化了过程复杂性和完成时间;但在cc阈值上限情况下,提出的方法检测精度比其他方法略差,因此提出的方法在适用性方面还有待进一步拓展。

参考文献(References)

[1] 张连成,魏强,唐秀存,等. 基于路径与端址跳变的SDN网络主动防御技术[J]. 计算机研究与发展, 2017, 54(12):2761-2771. (ZHANG L C, WEI Q, TANG X C, et al. Path and port address hopping based SDN proactive defense technology[J]. Journal of Computer Research and Development, 2017, 54(12): 2761-2771.)

[2] 刘江,张红旗,杨英杰,等. 基于主机安全状态迁移模型的动态网络防御有效性评估[J]. 电子与信息学报, 2017, 39(3):509-517. (LIU J, ZHANG H Q, YANG Y J, et al. Effectiveness evaluation of dynamic network defense based on host security state migration model[J]. Journal of Electronics and Information Technology, 2017, 39(3): 509-517.)

[3] HODO E, BELLEKENS X, HAMILTON A, et al. Threat analysis of IoT networks using artificial neural network intrusion detection system[C]// Proceedings of the 2016 International Symposium on Networks, Computers and Communications. Piscataway: IEEE, 2016:1-6.

[4] MONDAEEV M, ANKER T, MEYOUHAS Y. Method and apparatus for deep packet inspection for network intrusion detection: US20080031130[P]. 2013-05-21.

[5] QU X, YANG L, GUO K, et al. A survey on the development of self-organizing maps for unsupervised intrusion detection[J/OL]. Mobile Networks and Applications [2019-11-10]. https://link.springer.com/article/10.1007%2Fs11036-019-01353-0.

[6] CHIEN C F, HUANG Y C, HU C H. A hybrid approach of data mining and genetic algorithms for rehabilitation scheduling[J]. International Journal of Manufacturing Technology and Management, 2009, 16(1/2):76-100.

[7] KHALVATI L, KESHTGARY M, RIKHTEGAR N. Intrusion detection based on a novel hybrid learning approach[J]. Journal of AI and Data Mining, 2018, 6(1): 157-162.

[8] WANG W, LIU J, PITSILIS G, et al. Abstracting massive data for lightweight intrusion detection in computer networks[J]. Information Sciences, 2018, 433/434: 417-430.

[9] SULTANA N, CHILAMKURTI N, PENG W, et al. Survey on SDN based network intrusion detection system using machine learning approaches[J]. Peer-to-Peer Networking and Applications, 2019, 12(2): 493-501.

[10] 李龙杰,于洋,白伸伸,等. 基于二次训练技术的入侵检测方法研究[J]. 北京理工大学学报, 2017, 37(12):1246-1252. (LI L J, YU Y, BAI S S, et al. Intrusion detection model based on double training technique[J]. Transactions of Beijing Institute of Technology, 2017, 37(12): 1246-1252.)

[11] GAO X, SUN Q, XU H. Multiple-rank supervised canonical correlation analysis for feature extraction, fusion and recognition[J]. Expert Systems with Applications, 2017, 84:171-185.

[12] 刘雪娟,袁家斌,操凤萍. 云计算环境下面向数据分布的*K*-means聚类算法[J]. 小型微型计算机系统, 2017, 38(4):712-715. (LIU X J, YUAN J B, CAO F P. Data distribution *K*-means clustering for cloud computing[J]. Journal of Chinese Computer Systems, 2017, 38(4): 712-715.)

[13] XU T, CHANG H D, LIU G, et al. Hierarchical *K*-means method for clustering large-scale advanced metering infrastructure data[J]. IEEE Transactions on Power Delivery, 2017, 32(2): 609-616.

[14] PARK S, KIM J. A study on risk index to analyze the impact of port scan and to detect slow port scan in network intrusion detection[J]. Advanced Science Letters, 2017, 23(10):10329-10336.

[15] KORITSAS S, HAGILIASSIS N, CUZZILLO C. The outcomes and impact scale - revised: the psychometric properties of a scale assessing the impact of service provision[J]. Journal of Intellectual Disability Research, 2017, 61(5):450-460.

[16] 张冰涛,王小鹏,王履程,等. 基于图论的MANET入侵检测方法[J]. 电子与信息学报, 2018, 40(6):1446-1452. (ZHANG B T, WANG X P, WANG L C, et al. Intrusion detection method for MANET based on graph theory[J]. Journal of Electronics and Information Technology, 2018, 40(6):1446-1452.)

[17] KHROMYKH S V, TSYGANKOV A A, BURMAKINA G N, et al. Mantle-crust interaction in petrogenesis of the gabbro-granite association in the Preobrazhenka intrusion, Eastern Kazakhstan[J]. Petrology, 2018, 26(4):368-388.

[18] 叶子维,郭渊博,王宸东,等. 攻击图技术应用研究综述[J]. 通信学报, 2017, 38(11):121-132. (YE Z W, GUO Y B, WANG C D, et al. Survey on application of attack graph technology[J]. Journal on Communications, 2017, 38(11): 121-132.)

[19] SHI W, LU C, YE Y, et al. Assessment of the impact of sea-level rise on steady-state seawater intrusion in a layered coastal aquifer[J]. Journal of Hydrology, 2018, 563:851-862.

This work is partially supported by the Youth Program of National Natural Science Foundation of China (61802272).

WANG Lei, born in 1987, M. S., research fellow. His research interests include cloud computing, computer network security protection.