基于改进 AE-CM 模型的未知应用层协议识别

马甜甜,洪 征,陈 乾
(陆军工程大学 指挥控制工程学院,江苏 南京 210014)

摘 要:现有的未知协议识别方法存在提取的特征不够充分、聚类分配不准确等问题,影响了协议识别结果的准确性。AE-CM(deep autoencoder with embedding clustering module)解决了当前深度聚类模型异步优化的问题,提高了聚类分配的精度。该文提出的 DAEC-NM 协议识别模型以 AE-CM 为基础,通过加入高维卷积、时序卷积网络以及调整多层感知机结构的方法,改进了 AE-CM 的特征提取部分。为了更全面地获取协议信息,DAEC-NM 通过邻居分支采集邻居样本,并分析邻居样本间的局部关联特征,从而增强原样本特征中重要特征对聚类分配的指导能力。最后,采用了注意力机制来分析特征的重要性,以此为聚类模块设置有效的初始权重,解决了聚类模块在模型更新过程中权重特征更新较慢的问题。实验结果表明,DAEC-NM 能够有效提高未知协议识别的准确性。

关键词:网络流量;未知协议识别;深度自编码器;高斯混合聚类;嵌入层;邻居特征加权
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2024)03-0118-07
doi:10.3969/j.issn.1673-629X.2024.03.018

Unknown Application Layer Protocol Recognition Method Based on Improved AE-CM

MA Tian-tian, HONG Zheng, CHEN Qian
(School of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210014, China)

Abstract: Existing unknown protocol recognition methods suffer from insufficient feature extraction ability and inaccurate clustering assignments, which affect the accuracy of recognition results. AE-CM(deep autoencoder with embedding clustering module) addresses the issue of asynchronous optimization in deep clustering models and improves the accuracy of clustering assignments. The proposed DAEC-NM is based on the AE-CM. The feature extraction part of the AE-CM is improved by introducing high-dimensional convolution, temporal convolution network, and adjusting the structure of multi-layer perceptron. To obtain more comprehensive protocol information, DAEC-NM collects neighbor samples through the neighbor model and analyzes the local correlation features to ensure the accuracy of clustering results. Finally, we use an attention mechanism to capture the importance of features, and set effective initial weights for the clustering module to resolve the slow update problem in the clustering module. Experimental results show the DAEC-NM can effectively improve the accuracy of unknown protocol recognition.

Key words: network traffic; unknown protocol recognition; deep autoencoder; GMM clustering; embedding; neighbor feature weighting

0 引 言

未知协议识别是指在网络通信中识别出无法被预先确定的、类型未知的协议的过程。未知协议识别方法依据协议特征对网络流量进行分类,有助于发现异常的网络通信活动[1]。提高未知应用层协议的识别能力,有利于安全高效地提供网络服务[2]。

基于深度聚类的未知协议识别方法[3]通过神经网络模型对网络流量进行特征提取,并进行聚类分配。该方法适用于不同协议的特征和流量分布。

与此同时,基于深度聚类的未知协议识别方法[4]主要存在以下问题:

(1)特征提取和聚类分配是相互独立的过程,聚类结果不能指导特征提取,导致聚类性能不佳。

(2)未知协议的特征不确定,仅从时间或空间的单一维度提取特征会造成特征不充分。

(3)嵌入式聚类分配模块对不同特征的影响程度不同[5],但在分配初始权重时采用随机或相等的权重值,模型需要多次更新,收敛速度较慢。

收稿日期:2023-03-21 **修回日期**:2023-07-22
基金项目:国家重点研发计划项目(2019YFB2101704)
作者简介:马甜甜(1998-),女,硕士,研究方向为网络流量分析和逆向工程;通信作者:洪 征(1979-),男,副教授,博士,研究方向为网络流量分析和漏洞挖掘。

AE-CM[6]在现有深度聚类模型的基础上设计了嵌入式聚类分配模块,克服了聚类分配模块对特征提取模块指导性不强的问题。该文以 AE-CM 为基础,提出了未知协议识别模型(DAEC-NM)。该文的主要研究工作如下:

(1)提出了一种新的未知协议识别模型,改进了 AE-CM,并将改进模型应用于未知协议识别。

(2)在特征提取模块中插入重新设计的神经网络模块,增强了模型对协议时空特征的提取能力。经过特征提取模块获取的丰富特征能够被用于指导聚类簇的产生。

(3)使用 Two-branch[7]中提出的邻居模型提高协议识别的准确性。使用邻居分支来捕获邻居样本的格式信息和关联特征,并根据邻居特征提高主分支中相关协议特征的权重。

(4)在聚类模块中引入了注意力评分机制。记录模型特征提取过程中的特征权重,并在样本聚类分配过程中为相关特征设置合理的初始权重,指导样本进入相应的聚类簇。

(5)实验结果表明,与现有的基于深度聚类的未知协议识别方法相比,DAEC-NM 在 ACC、ARI 和 NMI 等指标上都有明显提升。

1　协议识别模型的设计

未知协议的识别主要包括协议数据预处理以及协议识别,如图 1 所示。数据预处理包括三个步骤。流量清理主要去除与协议识别无关的数据包,提高协议识别准确性。流重组和分割将网络流量转换为符合深度自编码器输入格式的数据,并将请求与响应组合在一起,便于分析协议内的关联关系。此后从网络流的开头截取固定长度的段,并根据需要执行截断和填充操作。最后,流量数据归一化对获得的固定长度的序列进行归一化操作,并将序列转换为固定格式的二维张量。通过数据预处理,可以提高输入数据的质量,减少数据噪声,保证模型训练的有效性和准确性。

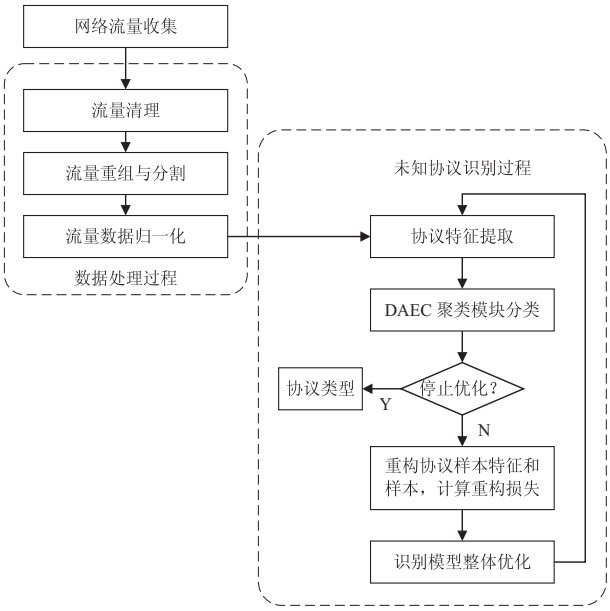

图 1　未知协议识别的流程

该文提出的 DAEC-NM 如图 2 所示,主要包含一个深度聚类分支(DAEC-branch)和一个邻居分支(NM-branch)。其中 DAEC 分支中包含协议特征的提取模块(feature extraction module)、聚类分配模块(clustering module)和协议重构模块(protocol reconstruction module)。各模块的具体工作将在后文详细论述。

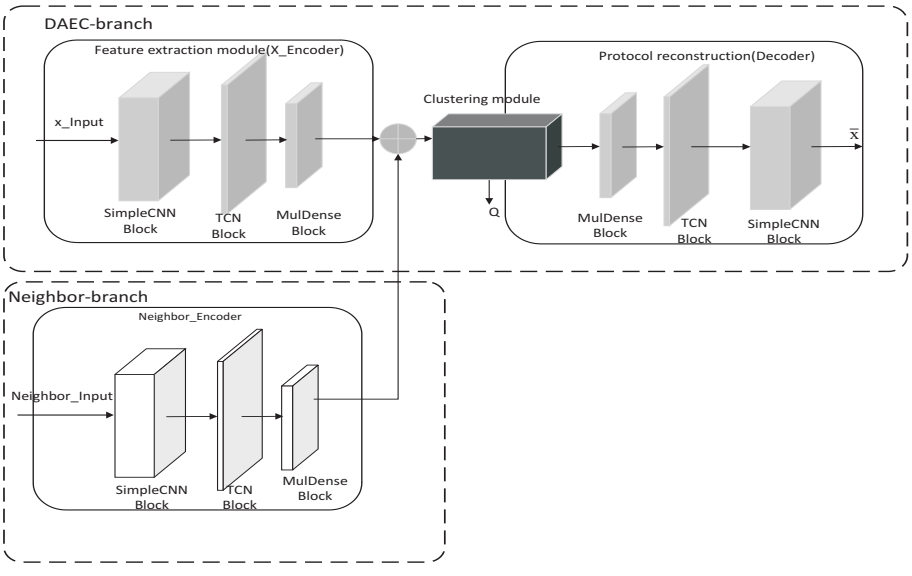

图 2　DAEC-NM 结构

1.1　协议识别模型的特征提取

深度聚类模型 DAEC 的设计如图 3 所示,特征提取编码器根据其功能模块将输入的数据流张量转换为

协议特征,聚类分配模块根据协议特征输出聚类簇分配结果,并根据协议簇分配结果重新调整协议特征权重,协议重构解码器根据协议特征重构协议样本。在设计未知协议识别模型中的特征提取编码器时,传统的堆栈自编码器处理非局部特征信息时不够灵活。DAEC 模型采用了简单卷积模块、时序卷积模块和多层感知机模块来提取协议样本的特征。简单卷积模块采用高维卷积对协议样本进行空间特征提取,以增强对不同协议的区分能力。时序卷积模块是一种时间序列模型,通过残差链接块和膨胀因果卷积提取协议数据中的时间相关特征,以提高聚类分配的准确性。多层感知机模块采用全连接层对前两个模块提取的特征进行组合和抽象,增强对协议特征的区分能力。为了避免模型计算冗余参数的压力,该文向 MLP 中插入衰减层 Dropout1,Dropout2,提高模型的泛化能力和稳健性。这些模块相互结合,可以挖掘协议数据的空间特征和时间特征,增强挖掘特征对不同协议的区分能力。

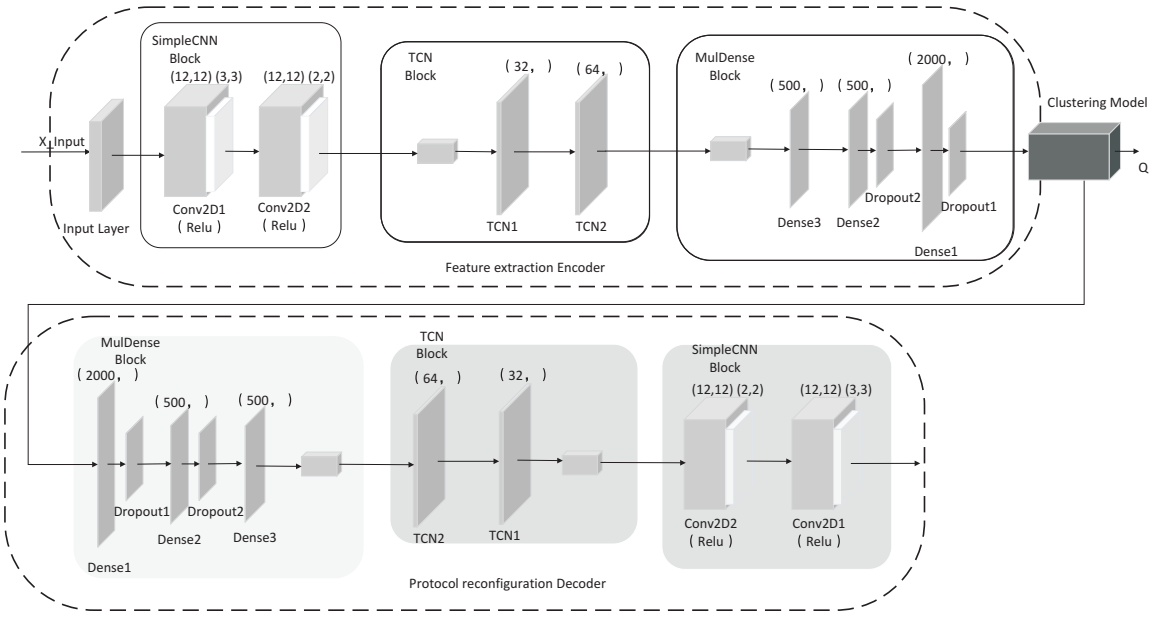

图 3 DAEC 自编码器模型结构

1.1.1 简单卷积模块

简单卷积模块由两个卷积层组成,用于提取协议样本的空间特征。协议特征在卷积层的输出值计算方法为:

$$y = \mathrm{ReLU}(W_y x + b_y) \tag{1}$$

其中,y 代表卷积操作后的输出,W_y 代表卷积核权重,b_y 表示偏置,ReLU 表示激活函数。

1.1.2 时序卷积模块

TCN 由多个残差模块(Residual Convolutional Block)构成,其处理函数为 TCN( ),它的正向传播计算过程如下:

$$h_t = \mathrm{ReLU}(\max(W_h x_t + b_h)) \tag{2}$$

$$n_t = \frac{W_n}{\| \sum W_n \|_2} * h_t \tag{3}$$

$$d_t = \begin{cases} n_t & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases} \tag{4}$$

$$f_t = W_f x_t + b_f \tag{5}$$

$$r = d_t' \oplus f_t \tag{6}$$

其中,h_t 表示经过膨胀因果卷积后的输出,第 i 个输入值 x_i 表示为 $x[i + j * r]$,j 表示当前卷积核的个数,r 表示膨胀率(dilation rate)。W_h,W_n 和 W_f 表示权重,b_h 和 b_f 表示偏置。n_t 表示归一化后的输出,$\| \sum W_n \|_2$ 表示权重的 L2 范数。d_t 表示衰退层后的输出,p 表示保留节点的概率,d_t' 表示 d_t 再次经过卷积层、归一化和衰退层后的输出。

1.1.3 多层感知机模块

从 TCN 模块输出的特征输入 MLP 模块,MLP 模块的正向传播计算公式如下:

$$z_1 = W_1 x + b_1 \tag{7}$$

$$d_1 = \frac{[\mathrm{rand}(z_1.\,\mathrm{shape}) < p_1]}{p_1} \tag{8}$$

$$x_1 = z_1 \cdot d_1 \tag{9}$$

$$z_2 = W_2 x_1 + b_2 \tag{10}$$

$$d_2 = \frac{[\mathrm{rand}(z_2.\,\mathrm{shape}) < p_2]}{p_2} \tag{11}$$

$$x_2 = z_2 \cdot d_2 \tag{12}$$

$$z_3 = W_3 x_2 + b_3 \tag{13}$$

其中,W_1,W_2 和 W_3 表示权重,b_1,b_2 和 b_3 表示偏置,p_1 和 p_2 表示衰退层保留节点的概率。

1.2 邻居分支与补充特征

高维特征空间表示可能无法捕捉到协议数据的语

义和局部联系。为了解决这些问题,该文采用了 Two-branch 方法设计邻居分支,采集邻居特征获得更丰富的协议信息,增强模型对协议的学习能力。邻居分支对邻居样本的二维张量进行特征提取,使用平均池化来得到一组邻居中心特征,从而获得一组通道特征和局部相关特征。这样,邻居分支可以帮助深度聚类模型更好地捕获协议数据的语义和局部联系,从而提高协议识别的准确性。

1.2.1　邻居编码器模块设计

模型的编码器可分为 X 编码器和邻居编码器,如图 4 所示。两个分支结构的网络模型相同且同步输入,X 编码器分支输入样本 x,邻居分支通过最近邻方法[8]将前次聚类结果中 x 的 k 个最近邻样本作为输入。假设一个输入邻居样本为 z,它的前向传播过程如下:

简单卷积模块的处理:

$$y_z = \mathrm{ReLU}(W_y z + b_y) \tag{14}$$

时序卷积模块的处理:

$$r_z = \mathrm{TCN}(y_z) \tag{15}$$

多层感知机模块的处理,简写为:

$$m_z = W_m z + b_m \tag{16}$$

其中,W_m 表示权重,b_m 表示偏置,需要通过多层感知机训练学习得到。

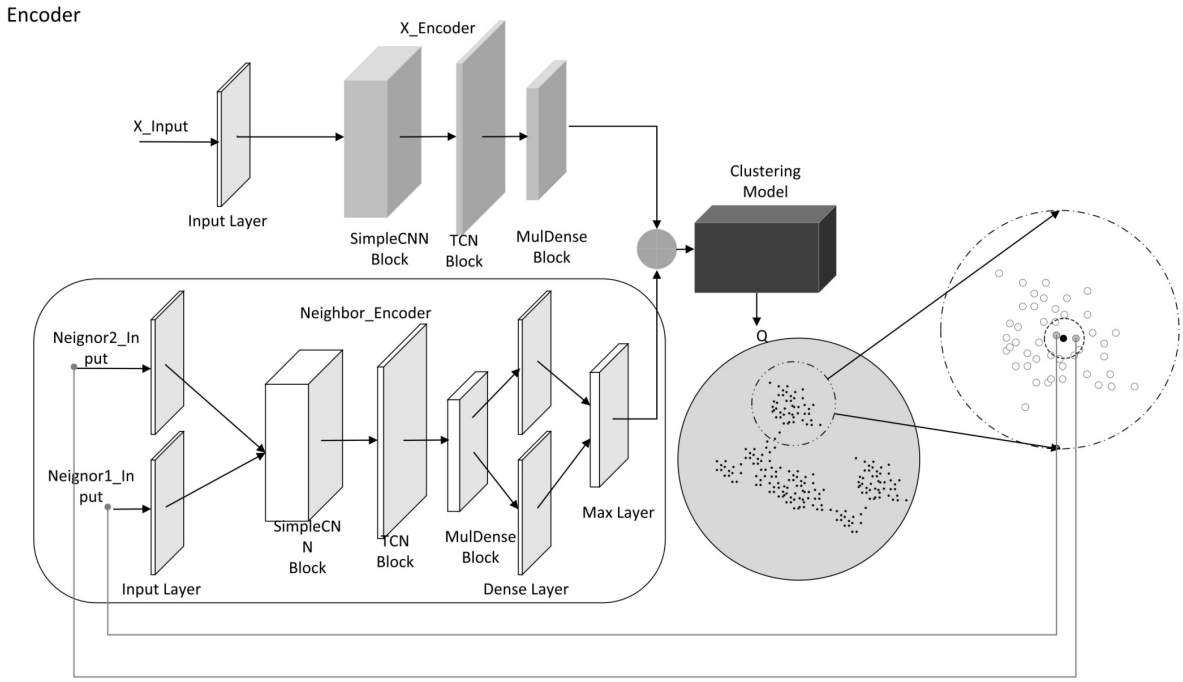

图 4　编码器模型邻居分支结构

1.2.2　邻居编码器权重加强设计

样本邻居的特征可以为协议分析提供更全面的信息,同时可以发现其他样本的特征模式,进而深入了解数据的总体特征。

X 分支和邻居分支获得整体样本特征表示 $\boldsymbol{z} = [z_1, z_e, z_3, \cdots, z_e]$ 和邻居特征表示 $\boldsymbol{m} = [m_{z1}, m_e, \cdots, m_{zk}, 0, 0, 0, \cdots](k < n)$,它们的特征权重分别是 W_z 和 W_m。为了增加整体样本特征中增强特征的权重,权重加强层设计如下:

$$W_z' = \theta W_m + (1 - \theta) W_z \tag{17}$$

其中,θ 表示增强参数,一般小于 0.5,防止邻居特征中的冗余特征过强,影响原样本特征。

将加强后的特征输入聚类层之前的隐藏层,筛选权重较强的特征,其前向传播计算为:

$$O_z = W_z' z + b_z' \tag{18}$$

1.3　基于注意力评分机制的聚类模块的设计

在 AE-CM 中使用 RMABs[9]映射方法,设计嵌入式聚类模块,它将原始特征空间映射到低维的嵌入空间,并在嵌入空间中进行聚类。

在 RMABs 映射方法中,将 EM 算法迭代过程封装为一个小型自编码器表示的聚类模块,映射过程如下:

$$\mathrm{F-step}: \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{K}}, F(x) = \langle p(z = k \mid x) \rangle_{\{1 \leqslant k \leqslant K\}} = \gamma \sim \mathrm{softmax}(XW_{\mathrm{enc}} + B_{\mathrm{enc}}) = \Gamma \tag{19}$$

$$\mathrm{G-step}: \mathbb{R}^{\mathrm{K}} \to \mathbb{R}^{\mathrm{d}}, G(\gamma) = \sum_{k=1}^{K} r_k \mu_k = \bar{x} \sim \Gamma W_{\mathrm{dec}} + B_{\mathrm{dec}} = \bar{X} \tag{20}$$

AE-CM 中聚类模块的 gamma 层的初始权重矩阵采用随机数进行设置,需要多轮训练来找到合适的特征权重初始值,降低了模型的识别效率,增加了训练开销。为解决这一问题,该文使用加性注意力评分机

制[10]生成初始权重矩阵。该评分机制捕获特征关联性和重要性评分,得到特征权重评分矩阵,并使用该评分矩阵作为初始化的 gamma 层权重矩阵,从而使 gamma 层在几次迭代后快速收敛。这种方法可以降低训练次数,减少训练开销,提高模型的识别效率,同时保持聚类模块的对称性。

聚类模块总体设计如图 5 所示,gamma 层对于上层网络提取的每一个协议特征,通过注意力评分函数获得注意力评分,用来表示该特征的重要性。在此基础上,对注意力评分进行加权平均 f,得到最终的初始注意力评分矩阵 $\boldsymbol{W}_h$,通过 gamma 层将该样本特征表示输入相应簇中。相较于公式 19,加入注意力评分机制的 gamma 层前向传播计算方法为:

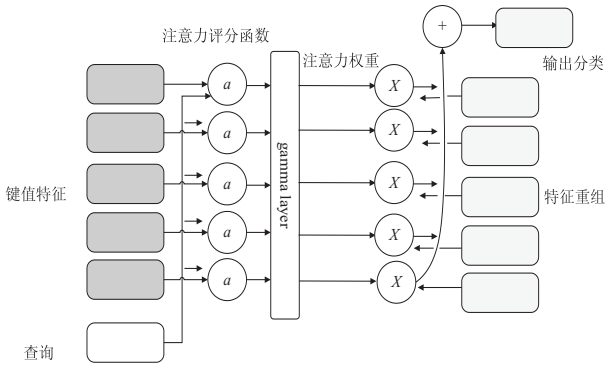

图 5 基于注意力评分机制的聚类模块

$$a = W_v^T \tanh(W_q q + W_h k) \tag{21}$$

$$h = W_h a + b_h \tag{22}$$

$$g = \mathrm{softmax}(h(q,k_i)) = \frac{\exp(h(q,k_i))}{\sum_{j=1}^{m}\exp(h(q,k_j))} = \Gamma \tag{23}$$

其中,W_v,W_q 和 W_k 表示权重,b_h 表示偏置,a 为注意力评分后加权平均层,h 为隐藏层,g 为 gamma 聚类层输出。

公式 20 中原 mu layer 是一个逆线性变化层 $\Gamma W_{\mathrm{dec}} + B_{\mathrm{dec}}$,为了抵消加性注意力机制的影响,聚类模块中 G-step 设计为一个多级权重衰减[11]的逆线性变化过程,mu layer 前向传播计算为:

$$\mathrm{mu} = \Gamma \bullet \mathrm{Dropout}(W_h) + b \tag{24}$$

其中,W_h 是 gamma 层提供的注意力权重。

1.4 协议识别过程

该文提出的未知协议识别模型,通过四个步骤(特征提取、聚类分配、协议重构、模型优化)进行协议识别。特征提取模块采用两个分支的设计,嵌入式聚类分配模块使用 gamma 层进行聚类分配,使用 mu 层调整未知协议特征的影响权重,最后使用重构误差和聚类损失来联合优化模型。模型不需要标记标签就能够自动识别未知协议的分布情况。

2 实验分析

基于 Tensoflow2.0 构建了未知协议识别原型系统。为评价模型的性能,主要考虑了以下标准:

(1)评价邻居数量以及邻居分支对协议识别性能的影响。

(2)评价改进聚类模块对协议识别性能的影响。

(3)与其他协议识别模型进行比较,并对协议识别模型的整体性能进行了评价。

2.1 数据集与评价标准

选择数据集 IDS2017[12]进行实验。数据集包含的是网络流量数据,以 pcap 的格式提供。提取了四种应用层协议(HTTP,FTP,DNS 和 SMB)进行测试。根据预处理方法对协议数据进行预处理后,得到了 45 514 条有序的网络流。在实验中,协议的标签被删除,从而使所有的协议都可以被视为未知的协议。为了评估所提出的协议识别模型的有效性,选择了精确度(ACC)、归一化互信息(NMI)和调整兰德指数(ARI)作为评估指标。

2.2 实验结果分析

2.2.1 邻居的影响

该文通过邻居分支提取样本邻居的特征作为补充特征。在邻居分支的设计过程中,通过近似算法选取 k 个样本邻居。需要设置实验分析邻居数量对协议识别效果是否有影响。此外,需要设置实验探究邻居分支对协议识别模型识别精度的提升效果。

(1)邻居数量的影响。

探究邻居数量对协议识别模型效果影响,设计 k 邻居数量范围为 1 ~ 300,实验通过 NMI 值随 k 值变化,探究邻居数量的影响。

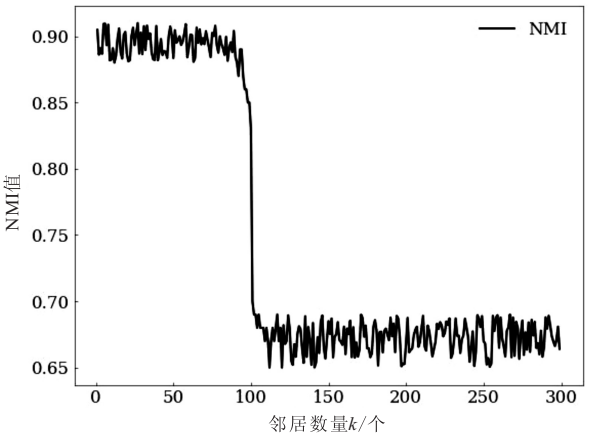

图 6 NMI 值随 k 的变化

从图 6 分析可得,邻居数量小于 90 时,NMI 曲线波动小,邻居数量超过 110 时,NMI 曲线恢复平稳波动;而邻居数量在 100 左右时,NMI 曲线急剧下降。

NMI 值在邻居数量小于 100 时较高(约 0.9),在邻居数量超过 100 时较低(约 0.66)。

从实验结果可以看出,较远处的邻居与原样本相差较大,从而导致原样本的特征被邻居样本的补充特征干扰。因此,该文选择较小数量的 k 值,能够呈现较好的协议识别效果。

(2)邻居分支的影响。

为探究邻居分支对协议识别模型的效果影响,通过实验比较带邻居分支的模型和不带分支的模型的识别效果。实验中分别获得协议识别模型(DAEC)带邻居分支(NM)和不带邻居分支(NA)情况下,模型的精确度(ACC)、调整兰德指数(ARI)和归一化互信息(NMI)。

表 1　有无邻居状态下协议识别模型表现　%

Model	ACC	ARI	NMI
DAEC-NM	98.04	96.03	92.82
DAEC_NA	97.28	94.91	91.52

从表 1 中得出以下结论:带邻居分支的深度聚类模型协议识别表现优于不带邻居的模型。

2.2.2　注意力评分机制在聚类模块中的影响

探究注意力评分机制在聚类模块中对协议识别模型的效果影响,通过实验比较聚类模块(CM)包含和不包含注意力评分机制对协议识别模型的效果。实验结果显示,相较于不包含注意力评分机制的聚类模块,包含注意力评分机制的聚类模块的协议识别模型表现更优。精确度提升了 4.23 百分点,ARI 指数提升了 11.63 百分点,NMI 指数提升了 10.78 百分点。其原因在于,注意力评分机制侧重于提高协议识别模型对于重要特征的关注度,有利于提高聚类效果。具体来说,注意力评分机制可以通过动态给不同的特征分配不同的权重,使得那些更具有代表性和区分度的特征能够更好地被聚类模块所利用。

表 2　有无注意力评分机制的聚类模块对协议识别模型表现的影响　%

Model	ACC	ARI	NMI
CM-attention score	98.04	96.03	92.82
CM-NA	93.81	84.40	82.04

2.2.3　与其他模型的横向比较

为了验证所提出的未知协议识别模型的性能,将该模型与 DEC[13], CAE[14], AE + K - Means[15], K - Means 和 GMM 进行横向比较。

其中,K-Means 和 GMM 是传统的聚类算法,常用于协议识别模型中。根据本研究实验结果设置它们的参数,K-Means 聚类簇数为 4,GMM 的成份数为 4。深度聚类的方法包含 DEC, CAE, AE + K - Means, AECM 以及 DAEC-NM(ours),其中除该文设计的网络外,其他方法的自编码器均为对称的堆栈多层感知机。

表 3　各种方法下协议识别模型表现　%

Model	ACC	ARI	NMI
K-Means	83.42	66.23	70.92
GMM	89.43	74.32	79.11
CAE	85.12	69.77	71.18
DEC	86.01	70.51	75.04
AE+K-means	84.86	70.51	74.08
DAEC-NM(ours)	98.04	96.03	92.82

不同协议识别模型的协议识别结果如表 3 所示。在表 3 所示的协议识别模型中,K-Means 和 GMM 表示基于机器学习的协议识别模型,其余为基于深度聚类的未知协议识别模型。实验结果表明,提出的深度聚类协议识别模型优于传统聚类模型,在此基础上,基于高斯混和聚类的模型优于基于 K-Means 聚类的模型。同时,嵌入式模型优于异步训练的模型,因为嵌入式模型能够更好地将聚类表现融入到编码器的训练中。此外,增加卷积模块的自编码器模型也优于原堆栈编码器模型,能够增强模型的时间、空间特征的提取能力,从而提高识别精度。总体而言,提出的协议识别模型比 DEC 模型在 ACC, ARI 和 NMI 评判标准上分别提高了 12.03 百分点,25.52 百分点和 17.78 百分点。

3 结束语

该文提出了一种未知应用层协议识别模型(DAEC-NM)。该模型的特征提取模块包含两个分支,主分支采用时空卷积网络来提取协议数据的时空特征,邻居分支捕获邻居样本间的局部关联特征作为补充。模型的聚类模块通过增加注意力评分机制的方法进一步优化识别模型,并实现聚类簇分配。实验结果表明,该模型在识别性能上优于其他协议识别模型。在未来的工作中,考虑把该模型应用于协议逆向分析、入侵检测等领域,为网络安全提供有效的保障。

参考文献:

[1] SHEIKH M S, PENG Y. Procedures, criteria, and machine learning techniques for network traffic classification: a survey [J]. IEEE Access, 2022, 10: 61135-61158.

[2] 马宝林. 未知应用层协议识别方法研究与系统实现[D]. 西安: 西安电子科技大学, 2021.

[3] 洪 征, 龚启缘, 冯文博, 等. 自适应聚类的未知应用层协议识别方法[J]. 计算机工程与应用, 2020, 56(5): 109-117.

[4] 冯文博, 洪 征, 吴礼发, 等. 基于卷积神经网络的应用层协议识别方法[J]. 计算机应用, 2019, 39(12): 3615-3621.

[5] PUNJ G, STEWART D W. Cluster analysis in marketing research: review and suggestions for application[J]. Journal of Marketing Research, 2012, 1983: 134-148.

[6] BOUBEKKI A, KAMPFFMEYER M, BREFELD U, et al. Joint optimization of an autoencoder for clustering and embedding[J]. Machine Learning, 2021, 110(7): 1901-1937.

[7] ZANGENEH E, RAHMATI M, MOHSENZADEH Y. Low resolution face recognition using a two-branch deep convolutional neural network architecture[J]. Expert Systems with Applications, 2020, 139: 112854.

[8] TANG S, YANG Y, MA Z, et al. Nearest neighborhood-based deep clustering for source data-absent unsupervised domain adaptation[J]. arXiv: 2107.12585, 2021.

[9] WANG Jinghua, JIANG Jianmin. Unsupervised deep clustering via adaptive GMM modeling and optimization[J]. Neurocomputing, 2021, 433: 199-211.

[10] WIEGREFFE S, PINTER Y. Attention is not not explanation [J]. arXiv: 1908.04626v2, 2019.

[11] CHEN W, SONG H. Automatic noise attenuation based on clustering and empirical wavelet transform[J]. Journal of Applied Geophysics, 2018, 159: 649-665.

[12] AGRAFIOTIS G, EFTYCHIA M, IOANNIS F, et al. Image-based neural network models for malware traffic classification using PCAP to picture conversion[C]//Proceedings of the 17th international conference on availability, reliability and security. Benevento: [s. n.], 2022: 1-7.

[13] GU C, WU W, SHI Y, et al. Method of unknown protocol classification based on autoencoder[J]. Journal on Communications, 2020, 41(6): 88-97.

[14] GUO L, WU Q, LIU S, et al. Deep learning-based real-time VPN encrypted traffic identification methods[J]. Journal of Real-Time Image Processing, 2020, 17: 103-114.

[15] FARD M M, THONET T, GAUSSIER E. Deep k-means: jointly clustering with k-means and learning representations [J]. Pattern Recognition Letters, 2020, 138: 185-195.