文章编号：2095-4980(2021)04-0712-05

一种基于改进自编码器的二进制协议聚类方法

彭博一，张 钊，蒋鸿宇*

(中国工程物理研究院 电子工程研究所，四川 绵阳 621999)

摘 要：针对无先验知识下，混合二进制协议数据帧难以识别分离的问题，提出了一种基于联合高斯混合模型(GMM)和自编码器的聚类方法。对于捕获到的未知二进制数据帧，首先通过栈式自编码器对其进行降维提取特征，并根据相应判别准则获取最佳聚类个数，最后使用改进了代价函数的自编码器对二进制数据帧进一步训练以提高聚类准确率。实验表明，该方法对网络二进制协议数据帧识别的准确率达到 94%以上。

关键词：二进制协议；降维；自编码器；特征聚类

中图分类号：TN915.07 **文献标志码：**A **doi：**10.11805/TKYDA2019556

A clustering method for binary protocol based on improved auto-encoder

PENG Boyi，ZHANG Zhao，JIANG Hongyu*

(Institute of Electronic Engineering，China Academy of Engineering Physics，Mianyang Sichuan 621999，China)

Abstract：To solve the difficulty of identifying the mixed binary protocol data frames without any prior knowledge，a clustering method based on joint Gaussian Mixture Model(GMM) and auto-encoder is proposed. For the captured unknown binary data frames，firstly its features are extracted via dimension reducing by stacked auto-encoder，and then the optimal number of clusters is obtained according to the corresponding criteria，finally the auto-encoder with modified cost function is utilized to train the binary data frames to improve clustering accuracy. The experimental results show that the accuracy of this method for recognizing the network binary protocol data frames is over 94%.

Keywords：binary protocol；dimension reduction；auto-encoder；feature clustering

在网络安全领域中，协议逆向工程在入侵检测、漏洞挖掘等方向都扮演着重要角色。而协议逆向工程的前提工作是对混合协议数据帧进行识别，其中二进制协议较之文本协议具有简洁紧凑、传输较快和易于机器读取的特点，因此被广泛应用。网络中公开的二进制协议，有已知的协议规范说明可以遵从，从而可以进行快速识别；然而，目前互联网中大部分私有协议的协议规范尚未公开，基于端口等传统方法并不适用，而人工方式[1]对协议数据进行解析往往耗时且易出错。因此，亟需新的方法能够快速准确地识别和分析网络中未知二进制协议数据帧。针对网络中未知二进制协议的识别问题，目前解决方法主要分为两种，分别是基于协议关键词的协议识别和基于协议统计特征的协议识别。王一鹏等人[2]提出应用 n-gram 对协议报文进行建模，然后基于面向支持向量机(Support Vector Machine，SVM)的主动学习方法进行协议识别；2015 年王勇等人[3]提出在多模式匹配中增加对二进制协议模式序列的数据统计，然后利用支持度和 FP-growth 算法获取特征串进行协议识别；2016 年罗建桢等人[4]提出将隐马尔科夫模型用于协议报文建模，然后基于最大似然概率确定协议关键词。以上研究基于协议关键词的在进行协议识别时，需要考虑到同一协议中不同类型具有相同协议关键词的区分，而且基于频率、长度等信息获取协议关键词存在一定的局限性。周洪川等人[5]提出基于改进的 K-means 方法对未知二进制协议报文进行聚类，但在聚类前需要人工进行预处理；Yan 等人[6]提出基于改进的主成分分析(Principal Component Analysis，PCA)和改进的密度峰值聚类(Density Peak Clustering，DPC)进行协议特征提取和聚类，但是降维特征提取只保证了原始数据的信息量，没有考虑所投影空间的方向；张路煜等人[7]提出基于卷积神经网络对网络协议进行识别，其局限性在于需要提供协议样本即测试集来训练神经网络，且在进行实验时，未知的协议的类型数量是给定的。

收稿日期：2019-12-19；**修回日期：**2020-02-05

*通信作者：蒋鸿宇 email:doherty2004@163.com

近年来，人工智能和深度学习等领域高速发展，其中用于无监督学习特征的自编码器[8]在图像处理、自然语言处理等领域中都取得了较好的效果。自编码器是一种期望输出等于输入的神经网络结构，通过最小化重构误差获取特征表示。在基于原始自编码器的基础上还发展了栈式自编码器、稀疏自编码器和去噪自编码器等。高斯混合模型(GMM)是一种典型的生成式模型，能够快速有效地处理大量训练数据，在提取特征后采用 GMM 模型进行后续聚类。本文面向网络中二进制协议数据帧，在无任何先验知识的情况下，提出了一种基于改进的栈式自编码器的协议聚类方法。考虑到二进制协议报文维度较高，需要进行降维以便于后续进行聚类识别，传统方法将降维与聚类两者分开来做，这样在无先验知识下提取的特征是随机的，需要人为进行修正，而本文提出的方法是将高维的二进制协议报文降维投影到对 GMM 聚类友好的低维空间上，以保证提取特征的完整与完备，同时提高聚类效率。

1　基于改进自编码器的聚类模型

1.1 模型介绍

本文研究的对象是从网络中捕获得到的混合二进制协议报文，二进制报文中存在大量无效数据，即无法据此确定数据类型，进而无法聚成簇。同时，二进制报文区别于文本类报文中以字节为单位的特性，二进制报文中域长度通常以比特为单位，且域与域之间没有如空格之类的分隔符，所以传统方法如 PI[9]、Discoverer[10]不适用。因此本文提出基于降维的报文聚类，即将混合二进制报文投影到低维度空间，获取最能表达报文内容的本征特征，并在低维空间完成聚类。整体架构如图 1 所示。方法步骤如下：

1) 特征提取。二进制协议报文维度较高，直接进行聚类往往难以确定报文的相似性，因此首先要进行降维，也就是特征提取。输入为二进制协议报文，通过自编码器模块，输出为降维数据特征。

2) 初始聚类及聚类簇数确定。没有任何先验知识，也就是无法知晓输入二进制报文的种类数。输入为降维数据特征，对此利用 GMM 进行初始聚类，并根据贝叶斯信息准则(Bayesian Information Criterion，BIC)，输出最佳聚类簇数。

3) 聚类优化。分开进行降维与聚类，导致两者之间缺乏联系，这里联合在一起构成改进的自编码器。输入为最佳聚类簇数和二进制协议报文，二进制协议报文通过改进的自编码器投影到给定空间，即有利于 GMM 聚类的低维空间，最后根据数据的低维映射进行聚类，输出报文最终聚类结果。

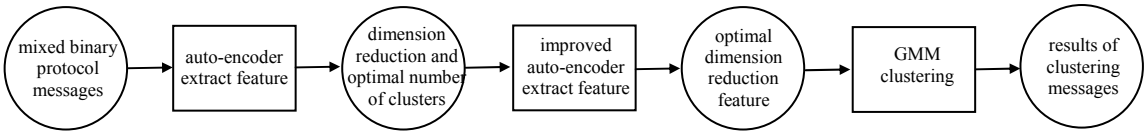

Fig.1 Clustering model of binary protocol messages
图 1　二进制协议报文聚类模型

1.2 特征提取

自编码器包括编码器和解码器，分别代表数据降维和数据重构，二者镜像对称，通过最小化重构原始数据的误差进行训练。当编码器输出维度小于原始维度时，可以获取原始数据中最显著的特征；而加深自编码器网络的层数即栈式自编码器则可以获取比浅层自编码器更好的压缩效率以及更强的表达能力。基本的栈式自编码器的网络结构如图 2 所示。

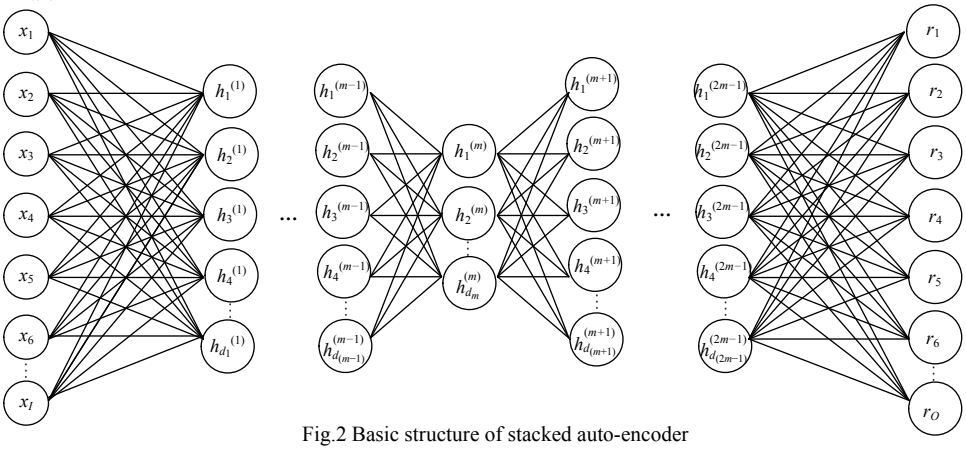

Fig.2 Basic structure of stacked auto-encoder
图 2　栈式自编码器的基本结构

栈自编码器中层与层的连接方式为全连接，第一层输出计算公式如式(1)所示：

$$\boldsymbol{h}^{(1)}=g^{(1)}(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}) \tag{1}$$

式中：$\boldsymbol{h}^{(1)}$为第一层输入层到隐含层的投影向量；$g^{(1)}$为第一层的激活函数；$\boldsymbol{W}^{(1)}$为第一层的权重矩阵；$\boldsymbol{x}$为输入向量；$\boldsymbol{b}^{(1)}$为第一层输出层的偏差向量。相似于式(1)，第二层输出为：

$$\boldsymbol{h}^{(2)}=g^{(2)}(\boldsymbol{W}^{(2)}\boldsymbol{h}^{(1)}+\boldsymbol{b}^{(2)}) \tag{2}$$

以此类推。其中栈式自编码器包括 1 层输入层，1 层输出层和 $2m-1$ 层隐藏层，输入层宽度和输出层宽度等于输入数据维度，隐藏层每一层的宽度$\{d_1,d_2,\cdots,d_{2m-1}\}$通过实验的训练得出。

1.3 初始聚类及聚类簇数确定

在得到二进制协议数据帧的降维表达之后，本文基于 GMM 进行聚类。GMM 是用概率密度函数来对一个特征数据集的统计分布进行描述。假设所有样本点来自 M 个参数不同的高斯分布，表达式如(3)所示：

$$p(x;\mu,\sigma^2)=\sum_{i=1}^{M}\pi_i N(x;\mu_i,\sigma_i^2) \tag{3}$$

式中：$N(x;\mu_i,\sigma_i^2)$为第 i 个高斯分布密度函数；π_i 为 $N(x;\mu_i,\sigma_i^2)$的权重，即归属该类数据的多少，$\pi_i>0$，$\sum\pi_i=1$，$N(x;\mu_i,\sigma_i^2)$表达式如(4)所示：

$$N\left(x;\mu_i,\sigma_i^2\right)=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{\left(x-\mu_i\right)^2}{2\sigma_i^2}\right) \tag{4}$$

式中：μ_i 和 σ_i^2 分别为第 i 个高斯分布密度函数中的均值与方差。求解 μ 和 σ^2 使得目标函数(5)最大：

$$P\left(X;\mu,\sigma^2\right)=\prod_{i}^{D}p\left(x_i;\mu,\sigma^2\right) \tag{5}$$

式中 D 为样本点的总数。最终得到每个样本点属于每个高斯分布即聚类簇的概率分布。然后，将每个样本点的概率分布中最大值所对应的簇标记为此样本点所属的簇。

理论上，聚类簇数应该等于 M(所有样本点来自 M 个参数不同的高斯分布)，且每个簇都唯一对应 M 中一个真实的高斯分布，然而真实的 M 值并不可知，还需确定最佳聚类簇数 M_b 作为真实的 M 的估计值。

先人为设定 M_b 可能的取值集合 S，如 $S=\{1,2,\cdots,9\}$，然后 S 中对每个可能的聚类簇数都进行一次聚类，在完成每个聚类簇数的聚类之后，利用 BIC 来确定最佳聚类簇数 M_b，计算方法如式(6)所示：

$$M_b=\arg\min_{k\in S}\left\{P_k\ln(k)-2\ln(L_k)\right\} \tag{6}$$

式中：P_k 为 GMM 中自由参量(均值、方差、权重)的总数；L_k 为 GMM 中的极大似然函数，L_k 的计算方法为：

$$L_k=\sum_{i=1}^{D}\log\left(\sum_{j=1}^{k}\exp\left(p_{ij}\right)\right) \tag{7}$$

式中 p_{ij} 的含义是对于样本 i 其归属于类别 j 的概率。选择使得 BIC 最小的聚类簇数 M_b 为之后进行协议聚类的真实类别数。

1.4 聚类优化

在得到二进制协议数据帧初始聚类概率分布之后，希望继续对其进行优化训练以提高准确率，即在基本的栈式自编码器的编码器输出添加聚类层 C，聚类层 C 是将自编码器中隐藏层输出即提取的特征作为 GMM 聚类的输入，如图 3。但是在无先验知识情况下，无法知晓各个样本的真实分布，而且自编码器经过训练进行降维所投影的空间是未知的，聚类的结果也会受到影响，因此本文提出通过联合自编码器重构误差代价函数和 GMM 聚类中极大似然函数来优化聚类概率分布，这样有目标地对原始样本进行降维到对 GMM 进行聚类友好的低维空间。聚类层 C 的目的是将降维所得到特征 $h^{(2)}$利用 GMM 进行聚类以得到最大似然函数 L，即式(7)，将 L 同重构误差相结合，以对二进制协议报文重新进行训练，改进后的代价函数计算方法如式(8)所示：

$$\min\sum_{i=1}^{D}\left(\left\|x_i-r_i\right\|_2^2+\frac{\lambda}{\log\left(\sum_{j=1}^{K}\exp\left(p_{ij}\right)\right)}\right) \tag{8}$$

式中：x_i 为第 i 个自编码器输入样本；r_i 为第 i 个自编码器输出样本；λ 为可调参数，起到平衡重构误差和 GMM 极大似然函数对总体误差的影响。

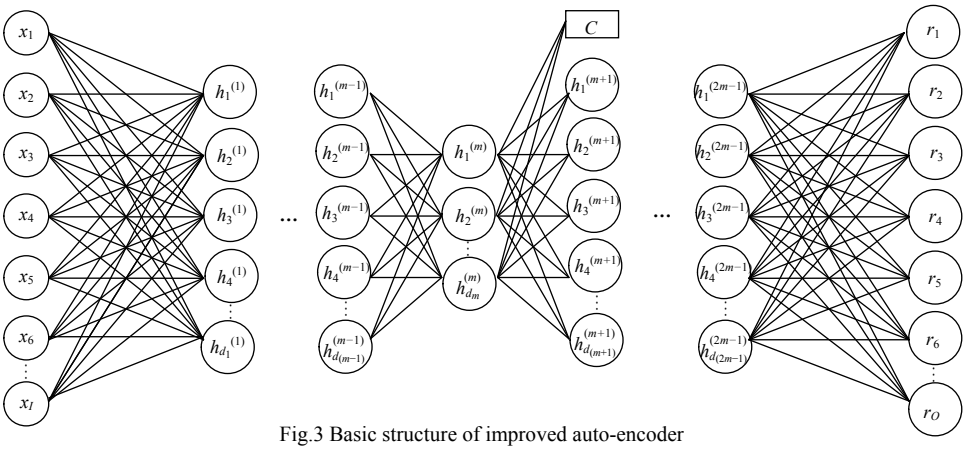

Fig.3 Basic structure of improved auto-encoder
图 3 改进自编码器的基本结构

2　实验结果与分析

2.1 数据集

本实验所采用的数据集是在真实网络环境中利用 Wireshark 所捕捉到的二进制协议报文，选取其中 5 种协议，包括 ARP,DNS,IGMPV2,CIFS/SMB,TLS，如表 1 所示，这些协议中同时包含定长协议与变长协议。

表 1 测试协议数据集
Table1 Data set of testing protocols

protocol	type	number
ARP	request	200
CIFS/SMB	–	200
DNS	request	200
IGMPv2	–	200
TLS	handshake	200

2.2 评价指标

利用匈牙利算法对二进制协议数据帧的聚类结果与其真实标签进行比对，得到最佳匹配。其中包括正确检测(True Positives，TP)，即聚类结果同真实标签一致的报文和错误检测(False Positives，FP)，即聚类结果同真实标签不同的报文。评价指标采用准确率，计算方法如式(9)所示：

$$prec = \frac{TP}{TP + FP} \tag{9}$$

2.3 实验结果

利用自编码器对二进制协议数据帧进行特征提取，并对降维之后的数据通过 GMM 进行聚类，得到具有不同中心数即聚类数的模型所对应的 BIC 分数，如图 4 所示。可以看出，当中心数为 5 时，BIC 分数最小，因此选用中心数为 5 的模型来进行聚类。

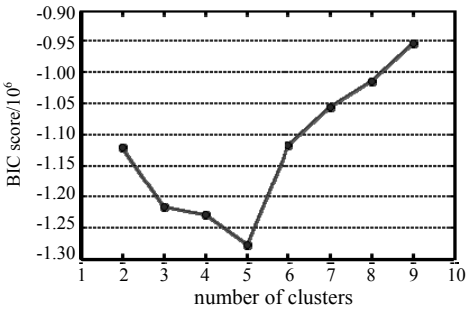

Fig.4 BIC scores under different clusters
图 4 不同聚类簇时的 BIC 分数

本文将自编码器和 GMM 联合起来聚类，并将聚类结果和真实结果相比照，最终得到 TP=946,FP=54，准确率达到 94.6%。

本文测试了单独使用改进 K-means[11]聚类方法、Yan 等人[6]提出的 PCA+DPC 聚类方法以同本文所使用的改进自编码器处理混合二进制协议数据帧进行对比，其中第一种方法类别数确定同样采用 BIC，第二种类别数确定采用该文提出的方法。表 2 是三种方法的准确率结果比对。

表 2 不同方法性能对比
Table2 Data set of testing protocols

method	accuracy
improved K-means	78.3%
PCA+DPC	82.3%
improved auto-encoder	94.6%

提出的基于改进自编码器的算法在处理混合二进制协议数据帧聚类的性能上表现良好。该算法能够在确定聚类簇数的同时，保证了混合二进制协议数据帧的降维表达投影到所需要的空间，即有利于 GMM 聚类的空间，完成了对于混合二进制协议数据帧的分离。经过网格搜索方法得到的局部最优自编码器的相关参数数据如表 3 所示。

表 3 改进自编码器相关参数列表
Table3 List of parameters used in improved auto-encoder

parameters	value
Λ	10
number of layers in neural network	7
width per layer	500/1000/200/20/200/1000/500

3 结论

协议识别是协议逆向工程的前提工作，在完成了协议识别的基础上，才能完成协议格式提取以及协议状态机提取。本文基于深度学习重新设计了以协议聚类为目的的神经网络结构，并基于真实网络环境提取 5 种协议进行验证。实验结果表明，本文提出的方法在无先验知识情况下对二进制协议报文进行聚类能够取得较好效果，相对传统方法提升了性能。

参考文献：

[1] TRIDGELL A. How Samba was written[R/OL]. [2020-02-05]. http://samba.org/ftp/tridge/misc/french_cafe.txt.

[2] 王一鹏,云晓春,张永铮,等. 基于主动学习和 SVM 方法的网络协议识别技术[J]. 通信学报, 2013,34(10):135-142. (WANG Yipeng,YUN Xiaochun,ZHANG Yongzheng,et al. Network protocol identification based on active learning and SVM algorithm[J]. Journal on Communications, 2013,34(10):135-142.)

[3] 王勇,吴艳梅,李芬,等. 面向比特流数据的未知协议关联分析与识别[J]. 计算机应用研究, 2015,32(1):243-248. (WANG Yong,WU Yanmei,LI Fen,et al. Protocol identification association analysis in mobile network environment[J]. Application Research of Computers, 2015,32(1):243-248.)

[4] 罗建桢,余顺争,蔡君. 基于最大似然概率的协议关键词长度确定方法[J]. 通信学报, 2016,37(6):119-128. (LUO Jianzhen,YU Shunzheng,CAI Jun. Method for determining the lengths of protocol keywords based on maximum likelihood probability[J]. Journal on Communications, 2016,37(6):119-128.)

[5] 周洪川. 面向比特流的未知协议分类研究[D]. 成都:电子科技大学, 2016. (ZHOU Hongchuan. Classification of bit-stream unknown protocol[D]. Chengdu,China: University of Electronic Science and Technology of China, 2016.)

[6] YAN X,LI Q,TAO S. A clustering algorithm for binary protocol data frames based on principal component analysis and density peaks clustering[C]// 2017 IEEE 17th International Conference on Communication Technology(ICCT). Chengdu, China:IEEE, 2017:1260-1266.

[7] 张路煜,廖鹏,赵俊峰,等. 基于卷积神经网络的未知协议识别方法[J]. 微电子学与计算机, 2018,35(7):134-136. (ZHANG Luyu,LIAO Peng,ZHAO Junfeng,et al. A method of unknown protocol recognition based on convolution neural network[J]. Microelectronics & Computer, 2018,35(7):134-136.)

[8] VINCENT P,LAROCHELLE H,LAJOIE I,et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010,11:3371-3408.

[9] BEDDOE Marshall. The protocol informatics project[R/OL]. Baseline Research, 2004. http://www.baselineresearch.net/PI.

[10] CUI W,KANNAN J,WANG H J. Discoverer: automatic protocol reverse engineering from network traces[C]// Proceedings of the 16th USENIX Security Symposium. Boston,MA,USA:[s.n.], 2007.

[11] 马俊宏,武丽芬. 一种改进的加速 K 均值聚类算法[J]. 太赫兹科学与电子信息学报, 2019,17(5):885-891,897. (MA Junhong,WU Lifen. An improved accelerated K means clustering algorithm[J]. Journal of Terahertz Science and Electronic Information Technology, 2019,17(5):885-891,897.)

作者简介：

彭博一(1995-)，男，硕士，研究实习员，主要研究方向为网络协议逆向. email:pengby253@163.com.

张 钊(1985-)，男，硕士，助理研究员，主要研究方向为网络协议逆向.

蒋鸿宇(1982-)，男，博士，研究员，主要研究方向为软件无线电与宽带数字接收机.