文章编号：2095-2163(2020)02-0355-05 中图分类号：TP391.4 文献标志码：A

基于半监督学习的网络应用流识别研究

赵 洋，余翔湛，郝科委
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：随着网络应用的发展普及，网络流量及网络规模迅速增大，产生的海量数据使得对网络应用流量的安全管理工作愈发艰难。传统的基于端口和载荷的应用流识别方法已经不能满足识别的精度要求。本文针对网络大量应用流识别问题，通过对现有少量标识数据的研究，采用半监督学习的方法提出并实现了无监督数据标识聚类，还采用有标识的方法进行辅助识别，可以为后续的监督学习提供大量的训练数据。

关键词：网络流量；半监督学习；无监督标识

Research on network application stream recognition based on semi-supervised learning

ZHAO Yang, YU Xiangzhan, HAO Kewei
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the development and popularization of network application, network traffic and network scale increase rapidly, resulting in massive data making the security management of network application traffic more and more difficult. The traditional application flow recognition method based on port and load can no longer meet the precision requirement of identification. In this paper, aiming at the problem of network application stream recognition, a semi-supervised learning method is proposed and implemented for unsupervised data identification clustering through the research of a small amount of existing identification data, and the labeled method is used to assist the identification. A large amount of training data can be provided for subsequent supervised learning.

【Key words】 network traffic; semi-supervised learning; unsupervised identification

0 引 言

目前，随着互联网技术的飞速发展和网络设施的不断升级进步，越来越多的网络应用已经进入人们日常生活中来。人们对互联网技术的认可度提升以及网络的应用范围也日趋宽广，使得人们的生活越来越依赖来自这些网络应用所提供的服务。网络应用在各领域的普及推广不但为人们的日常生活带来便利，同时更极大提高了社会的工作效率。目前，宽带计入能力的提升、不断更新的通信方式、“三网融合”工程的加速开展、“百兆乡村”政策的出台、物联网技术的应用与发展以及“互联网+”重大工程的实施，综上的论述都切实表明中国正处在、并将会长期处在全民网络时代。但随着网络应用的强劲拓展态势，网络流量及网络规模迅速增大，产生的海量数据使得对网络应用流量的安全管理工作愈发艰难。同时，由于互联网的虚拟性、开放性和交互性，使得网络应用质量参差不齐，良莠混杂，甚至还有某些不良网络应用利用现在先进的技术，假借正常端口或者协议来传播。而且，悄然伺机而动的病毒、木马也会伴随着新的网络应用，威胁着用户的隐私数据安全，给人们带来巨大的损失。因此，随着国内计算机技术的广泛应用与飞速发展，网络安全已跃升至国家安全战略地位，“没有网络安全就没有国家安全”的理念已日益深入人心。作为网络安全的重要环节，网络应用识别技术的研究尤为重要。

1 研究现状

目前识别方法主要分为机器学习识别和非机器学习模型两种。其中，非机器包括基于端口的报文识别检测技术和基于负载的识别检测技术，随着网络应用技术成果的相继问世，这些传统的方法已经难以适应不断变化的协议规则，因而逐渐为机器学习方法所取代。

相对于非机器学习，机器学习方法更加依赖于数据包和数据流特征而不是简单的特殊字段识别和

基金项目：国家自然科学基金(61771166)；国家重点研发计划项目(2016QY05X1000)。

作者简介：赵 洋(1994-)，男，硕士研究生，主要研究方向：信息安全、应用流识别；余翔湛(1973-)，男，博士，教授，博士生导师，主要研究方向：信息内容安全、网络安全、物联网安全等；郝科委(1993-)，男，硕士研究生，主要研究方向：信息网络、网络安全。

收稿日期：2018-06-06

匹配。影响机器学习方法主要取决于 2 个方面:特征提取方法和分类算法选择。其中,特征选择可以定制在 2 个层面上:数据包和数据流。数据包特征是通过对数据流一定范围内数据包的特征,诸如:最长包长、最短包长、平均包长、包长中位数方差等信息进行统计,最终整合得到结论。数据流特征则是包括:客户端端口、服务器端端口、数据流平均包长、数据流空包数、数据包传输平均时间间隔等特征,对应用流或是应用进行识别。一个好的识别模型,一般都会根据所识别的内容特性,选用两者中的适当内容进行分类训练识别。这里,对目前主流的研究方法可阐释论述如下。

(1)基于端口的网络应用识别。这是人们最早用来识别网络数据流路的方法。在早期的简单网络中,网络应用种类少且大都使用特殊的端口号,所以只需要观察并识别传输层报文头中的端口号,就可以辨识出相应的网络应用。这种识别方法不仅高效,而且所耗费的资源也是所有方法最低。起初,大部分网络都会选择特定端口号,而且不同种类的应用一般都配有不同的传输端口。基于端口号的应用识别技术便可以根据人工统计,选择特定的报文传输端口来确定目前应用类型。

(2)基于载荷的应用流识别方法。这是基于端口识别方法的传承和进化。相对于基于端口的识别方法,基于载荷的识别方法选择了应用层数据中的特殊字段,通过对大量的应用层协议的分析和统计,找出属于每一个应用层协议的特征码,再通过新来的数据流与特征码的整合匹配,得出识别效果。考虑到每种协议都具有其特定的规则和使用方式,所以一个好的特征码提取算法和特征码匹配算法往往会取得非常好的识别准确率和效率。

(3)决策树。是在已知标签数据分析基础上,通过构建决策树来求取净现值的期望值大于等于零的概率,判断可行性的一种决策方法。在众多数据挖掘和机器学习研究中,决策树归纳法是应用最广的方法之一。决策树中的每个节点代表在一个识别过程中的测试或是识别,若其含有分支则表示当前节点的识别结果,每个叶节点代表其最后的类型。

(4)基于神经网络方法。分析可知,数据量较少的时候,决策树的准确率、效率都优于神经网络。但随着训练数据的不断增加,学习强度的不断上升,神经网络的准确性能将更加出色。特别是随着新应用的渐次出现,一些直观的属性已经难以完全区分应用类型,特征的选取也越发困难,那么基于神经网络的识别方法在应用识别方面就尤为突显其强大适用性了。

根据准确性、复杂性、拓展性以及加密流量识别能力,本文对上述 4 种方法进行了对比分析,得到的结果见表 1。

表 1 网络应用识别的方法对比

Tab. 1 Comparison of methods of network application identification

	准确性	复杂度	拓展性	加密流量识别
端口识别法	低	低	低	否
载荷识别法	一般	低	低	否
特征识别法	一般	一般	一般	能
神经网络	高	高	高	能

2 数据聚类标识法

2.1 输入数据流特征选择

伴随着人工智能的发展,机器学习的方法已经成为各领域解决问题的重要方法。基于决策树、行为特征的方法都使得应用流识别的准确率大大提升。传统的基于机器学习的应用流识别方法一般是基于以往训练经验,选择最具有代表性的数据包或数据流的具体特征集,通过对特征集合向量化作为训练模型和测试部分的输入。而后,即是不断调整决策树构造或者隐藏层的权值,使训练集识别准确率达到最优。但是目前的研究现状是,网络应用及网络协议的数量已经越来越多,有限数量的显示特征已经不能完全地作为当前网络流量的代表集合,自动去寻找代表特征集合就非常重要。而且现实场景中对网络数据流的标识工程量较大、难度高,所以应用与训练的标识数据相对于海量的未标识数据少之又少,基于监督学习的神经网络分类器很难直接从少量的标识数据流中学得准确识别信息。因此融合了监督学习与无监督学习优势的半监督学习方法随即提出了通过在无监督学习提供大量标识数据的基础上,再使用监督学习建立分类器的方法。

2.1.1 基于五元组的数据流拼接

本文训练数据来自由抓包软件从网卡抓取的离线 pcap 文件,测试数据则是实时从网卡抓取的 pcap 文件。由于实际网络中会含有多进程通信,应用流自然不会单独出现在实际网络中,为了更好地识别网络应用流种类,就需要将网络应用流拼接起来。

在实际的网络应用流分析中,研究发现多数情况下,在一个较短的时间内同样的 2 个 ip 使用相同端口一般只通信一次。这使得可以通过对五元组的组合计算,将同一个短期的 pcap 中所有应用流单独拼接出来,并按顺序存成对应数据。在本研究课题

中，每个数据流的区别特征值 *key* 的计算公式为：

$$key = str(ip.src) + str(ip.dst) + str(port.src) + str(port.dst). \quad (1)$$

在一个 pcap 中，每读入一个数据包，将计算其 *key* 值，根据 *key* 值将相同的数据包拼接在一起。数据流的拼接过程详见图 1。

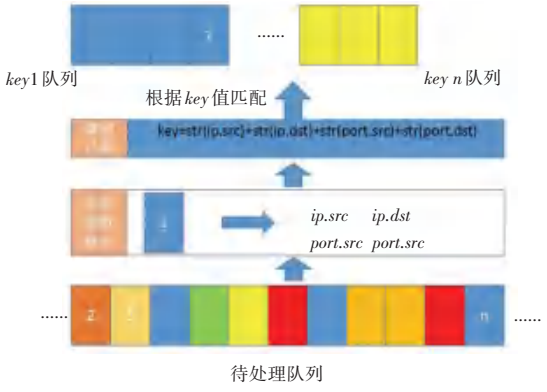


图 1　pcap 数据流拼接

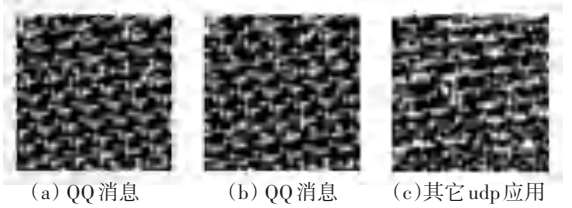
Fig. 1　pcap data stream splicing

2.1.2　数据特征选择

根据数据流拼接的结果，研究获得了单个数据流信息。但是在实际网络中，将很难保证每个抓取的数据流都不存在缺失、重传、或是截取不完整等问题，同时很多基于流的特征并不能作为最好的选择去代替这个应用流，而基于数据包的特征也不能去代替整个应用流。

在解决数据流代表性问题上，本文使用了数据流原文作为训练的输入。通过观察发现，数据流原文是一串十六进制的数字，而 2 个十六进制的数字则最终组成了 0~255 的数字，并且恰好对应了灰度图像中的灰度值范围，使用深度学习的研究思路也随即广受关注。而且，由于每个流的长度不同，数据包个数、甚至每个数据包大小也不同，就需要选取每个数据流的相同数量、长度的报文作为特征向量。

首先将每种应用报文按照一字节 8 位为一维特征，将每种应用的应用流拼接成图像，通过对不同类型的数据流图像进行对比，如图 2 所示，发现相同的应用类型，如图 2(a)与(b)均为 QQ 消息数据流，具有相似的图像；而不同的应用类型的数据流原报文图像则如 2(c)所示，与前 2 个 QQ 图像存在较大的差异，所以使用原报文方法是可行的。而后，根据文献[1-4]识别研究过程的原理解析，研究分别选择包长、数据包应用层协议类型、数据包数据段长度等显性特征来绘制出图像；并对 TCP 头设置 push 位包数、从客户端到服务器方向，以初始端口发送 tcp 负载大小和从服务器到客户端平均负载大小等基于数据流的特征进行统计分析。图 3 随即展示了 QQ 聊天与其他 udp 应用前 50 数据包长度统计对比。其中，蓝色和绿色的线条代表 QQ 聊天，橙色代表其它的 udp 应用。显而易见，在前 50 数据包长度对比上，相似的应用同样具有相似的性质。与此同时，研究还针对其它特征都进行了比对，效果大致相似。



(a) QQ 消息数据流 1　(b) QQ 消息数据流 2　(c)其它 udp 应用图像

(a) QQ message data stream1　(b) QQ message data stream2　(c) Other udp application feature image

图 2　QQ 聊天与其它 udp 应用的特征图片对比

Fig. 2　QQ chat with other udp applications

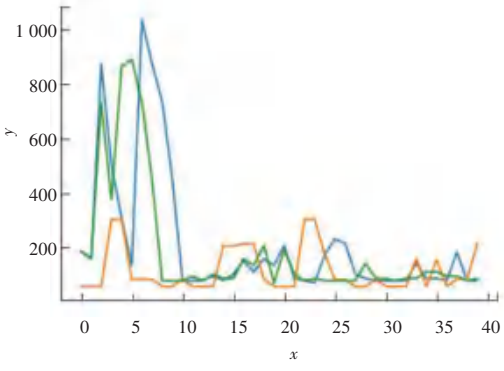


图 3　QQ 聊天与其它 udp 应用前 50 数据包长度统计

Fig. 3　Top 50 packet length statistics for QQ chat and other udp applications

为此，可推得如下研究结论：每个流前 50~100 报文由于其包含应用流建立连接和控制报文的交换信息，而且也会带有少量的其它通信信息，故而选择前 50 个数据包能够有效地代表数据流。而在每个数据包中，使用相同的传输层协议往往具有相似的传输层结构，不能很好地代表报文特征。研究中为区分应用流，则选择使用了应用层报文。通过统计分析，选择前 50 字节作为每个数据包的代表特征值。这样一来，每个数据流就可以使用 50 * 50 = 2 500维数据作为输入向量训练模型。

2.2　基于自编码的数据降维

在聚类的开始阶段，通过分析观察报文的原文则会发现，有很多的报文原文中数值为 0，且数据段相对较短的报文内容向量，本文也对其进行了补 0 处理，这里为了使距离度量相似性的设定不会失效，将首先使用数据降维的方法对输入的矩阵向量做出

降维处理。

与传统识别方法提取数据流、数据包特征识别方法不同,基于数据包原文的识别方法在每个维度上取值范围、代表含义都是相同的。这使得在维度下降方面,基于报文原文的方法可以使用相对优质的特征下降法而不仅局限于特征选择。通过试验对比分析,研究选择使用自编码器降维方式。对此,文中将给出研究论述如下。

2.2.1 自编码器模型

AutoEncoder 是一个将数据的高维特征进行压缩降维编码,再经过相反解码过程的一种学习方法。学习过程中通过解码得到的最终结果与原数据进行比较,再根据修正权重偏置参数降低损失函数,不断提高对原数据的复原能力。学习结束后,前半段的编码过程得到结果即可代表原数据的低维“特征值”。通过学习得到的自编码器模型可以实现将高维数据压缩至所期望的维度,原理与 PCA 相似。本课题使用的自编码器结构则如图 4 所示。输入是由每个数据流前 50 数据包,每个数据包使用前 50 字节,共2 500维向量组成。中间通过对隐藏层的训练,选择最优的隐藏层权值,使得还原结果更加准确,也就是说使得输出层的低维向量更具有代表性。

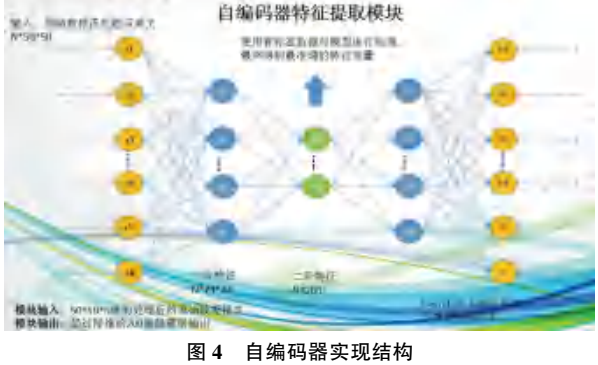


图 4 自编码器实现结构

Fig. 4 Self-encoder implementation structure

2.2.2 自编码维度选择

选择不同的维度对原始数据进行表达会产生不同的表达效果。为了使自编码器能够对原始数据的表达性更强,研究分别将输出层数设置为 10~600,并将 2 万多组的网络数据流分为 20 组,对每组均采用了编码/解码操作,通过求取 20 组平均前后数据方差值,描绘后的展现即如图 5 所示。由结果显示可知,选取 200 维作为最终聚类维数不但降维效果很好,而且还原度也相对较高。

2.3 基于 k-means 的数据聚类标识法

在降维后,数据变为 200 维特征的矩阵集。为了能够获得充足的标识数据作为构造分类器的训练数据,半监督分类方法选择使用无监督聚类结合少量标签数据对大量的未标识数据进行标识操作。在聚类方法选择上,根据目前半监督分类和聚类应用于数据流识别的现状,研究选择聚类效果较好的k-means算法进行聚类标识。使用k-means聚类标识数据的研发过程详述如下。

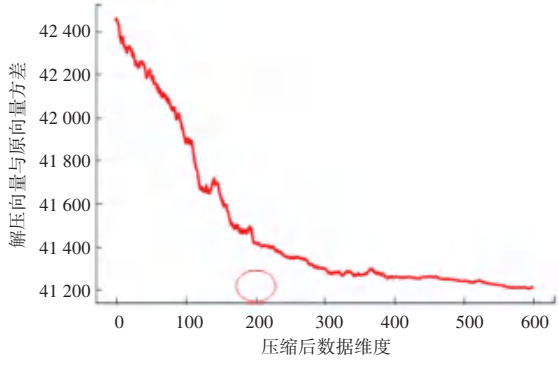


图 5 压缩维度选择与还原效果关系图

Fig. 5 Compressed dimension selection and restore effect diagram

2.3.1 k-means 算法 k 值选择

k 值作为 k - means 算法的核心关键点之一,其选择策略对于最终聚类效果有着至关重要的影响。与传统的k - means 以中心点收敛为终止条件不同,由于聚类的数据流存在新类别,使得中心点应当具备一定的数量调整能力。基于此,本文使用了循环聚类的方法,将每次的聚类结果作为下一次聚类方法选择的判定条件。以一定的距离作为阈值,阈值之外的点作为本轮未标注点,如果未标注点达到一定数量,则启用 $k+1$ 作为下轮 k - means 的 k 值,重新选择中心聚类,直至 k 值不变并收敛。如果中心点收敛且未标识数据没有达到阈值,聚类结束。

2.3.2 k - means 算法距离选择

传统 k - means 算法一般以欧式距离为衡量类别间相似的标准,但对于数据包原文来说,虽然每一位的取值范围相同,但每一维度所代表含义的差异可能使传统欧氏距离的区分效果大打折扣。本文选择加权的欧氏距离作为各点之间的距离度量方法,可以避免维度特征之间的差异。

加权的欧氏距离也可以解读为标准化欧氏距离,是针对欧氏距离的一种改进,在计算距离前将对每一个维度进行数据标准化使得期望为 0,方差为 1。先求出带标识数据在第 n 维度上的标准差 s_n,对于 2 个向量 $\boldsymbol{a}(x_1,x_2\ldots,x_k)$ 和向量 $\boldsymbol{b}(y_1,y_2,\ldots,y_k)$ 之间的加权距离公式可以表示为:

$$d_{ab} = \sqrt{\sum_{n=1}^{k} \left(\frac{x_n - y_n}{s_n} \right)^2} . \qquad (2)$$

2.3.3　k-means 算法实现描述

算法　基于加权欧氏距离的 k-means 算法

输入：带标签和未带标签的 2 组数据集

输出：识别之后带标签数据集及标签集

Step 1　通过对有标识数据的统计，得出现有标识类别数作为 k 初始值。

Step 2　将带标识数据按照标签分别存入不同的集合中。

Step 3　计算所有数据在各个维度上的标准差。

Step 4　分别计算各个集合标签集中向量在各维度上的均值，组成各个集合的初始 k 个中心点。

Step 5　分别计算各个集合中距离中心点最远的距离作为本轮阈值 d。

Step 6　带标记的向量不动，分别计算不带标记向量到各个中心加权距离，如果该点所有中心点最小距离大于 d，则将该数据暂时放入 *unknow* 队列。如果最小距离小于 d，则将其归入距离最近的集合中。

Step 7　将所有新集合向量各个维度取均值作为新的中心点，若中心点与上轮不同，重复 Step 6。

Step 8　如果中心点相同，统计 *unknow* 数量，若大于本次聚类标签数最少的类别数，则将 $k + 1$，取 *unknow* 数组中位数下标的向量作为新的聚类中心，重新进入 Step 6。

Step 9　若小于最少类别数，则将 *unknow* 数据抛弃。对当前每个集合中的数据进行分组标记。对于新分出来的集合采用人工标记法，随机抽取一定数量的应用流进行人工识别，对标识结果进行比对。若最多类型数量超过 90%，使用该类型标识这个集合，否则舍弃。

3　实验结果与分析

通过对 32 000 组标记数据流进行模拟，并选择分组聚类标识法测试，其中包括 coco 数据流 8 378 条，zello 数据流 7 693 条，skype 数据流 7 752 条，ftp 站点数据流 3 653 条，随机应用流 4 524 条。选择 4 种有标记数据流各 1 000 条作为已标识数据集。其余的 28 000 条以 4 000 为一组作为未标识应用集。使用已知标识的数据集分别与每组未知标识数据进行聚类标记，通过与原标记进行对比识别，得识别运行结果详见表 2。

表 2　聚类结果统计

Tab. 2　Statistics of clustering results

	正确	错误	丢弃	其它
1 组	3 357	123	15	505
2 组	3 412	150	32	406
3 组	3 103	110	14	773
4 组	3 347	87	12	554
5 组	3 240	121	11	628
6 组	3 308	97	21	574
7 组	3 052	112	17	819
总计	22 819	800	122	4 259

接下来在表 2 基础上，处理得出识别准确率的仿真运行结果，如图 6 所示。

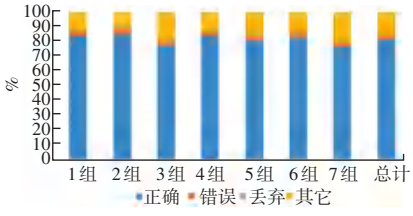


图 6　聚类准确率的运算结果

Fig. 6　The results of clustering accuracy

图 6 给出的聚类后根据识别效果对每组识别准确率进行统计显示，每组标识数据识别错误率均不超过 5%，因距离过远而丢弃的数据都不足 1%，而标识为其它的数据与已知的未标注数据在总量上彼此相近。结合在一起，可以判定总体数据的聚类标识准确度达到 95%以上，该效果可以用来对未标识数据进行有效的标注。

4　结束语

针对不断出现的新应用流的识别，传统的非机器学习方法无法对新类型应用进行识别，只能够重新建立模型；而传统的基于特征的机器学习方法也很容易出现识别错误和特征选择不具典型性的问题。基于数据报文的应用流识别使得识别过程可以从应用流本身挖掘特征而非仅依赖于选择的特定特征，极大地增强了模型的自身学习能力和对新应用类型识别和学习的适应性。在分类识别之前，大量的有标识应用流是必要的，通过半监督学习的方式可以采用少量的标识数据对大量的未标识数据进行准确的标记，从而为准确的监督学习模型分类器的建立提供坚实的基础。

参考文献

[1] 史可. 高性能网络应用协议识别技术的研究与应用[D]. 北京：北京邮电大学，2015.

（下转第 364 页）