异常检测数据集与属性调研

作者: 李侗格

一、异常检测概述

异常(Anomaly)，是指目标数据集中与正常模式的统计特征明显不一致的数据模式，在很多领域的数据分析中都是一个比较常见的现象。造成数据异常的活动多种多样，比如金融数据中的信用卡欺诈，网络数据中的网络攻击等。

按照数据具体表现形式，异常可以分为以下三：点异常(Pointanomaly)，指某项单独的数据与正常模式差异程度较大，例如，工业数据中，如果某辆车平时的耗油量都是每天五升，但是有一天这个数值变成了五十升，那么这项数据就是点异常；上下文异常(Contextualanomaly)，指某项数据在特定上下文环境中与正常模式差异较大，例如，金融数据中，信用卡消费记录会在某些节日期间开支变大，例如圣诞节或新年，这属于正常现象，但是如果在非节假日期间某人的信用卡开支突然变大，这就属于上下文异常；集合异常(Collectiveanomalv)，指一组相似数据共同表现出与整个数据集的正常模式差异较大，例如，医疗数据中，如果某人的心电图出现一段较长时间的连续低值，则很有可能是一种叫做期外收缩的心脏异常现象造成的，这属于集合异常，相反地，单独的一个低值并不属于异常。

引起网络流量异常的原因有操作不当和遭受网络攻击两种。本文主要关注由网络攻击造成的网络流量异常现象。网络攻击，是指任何威胁到计算机网络系统安全正常运行的有特定目的的行为。从形式上分为四类。

1. 拒绝服务(DoS，DenialofService)：在DoS攻击中，系统被大量来自单个连接的请求充斥，直至所有资源被耗尽，然后再也无法处理合法请求。在分布式拒绝服务攻击(DDoS)中攻击者使用分布在互联网上的多个连接且可能是僵尸网络的一部分，这种攻击使得用户无法使用服务或资源。小，中型DoS和DDoS攻击通常被攻击者当作幌子，来隐藏更小但更危险的活动或拿下安全设备如防火墙。
2. 窃听(Probe)：这种攻击的目的是收集目标主机或者网络的信息，本质上是一种侦查活动。对于一个网络的探测攻击可以获知其主机的数量和类型，对于一个主机的探测攻击可以获知其安装的应用程序和类型。探测攻击本身并不会造成实际的损失，但是探测攻击往往是实施进一步网络攻击的第一个步骤，利用收集的目标信息可以实施更加准确和有效的攻击活动。
3. 用户到根(U2R，UsertoRoot)：在这种攻击时，攻击者通常已具有目标系统的本地访问权限，意图利用由普通用户到超级用户/管理员的升级权限漏洞。最常见的U2R攻击是缓冲区溢出(bufferoverflow)攻击，攻击者试图溢出缓冲区以提升其用户权限并执行恶意代码。
4. 远程到本地(R2L，RemotetoLocal)：在此类攻击中，攻击者在目标计算机中没有账户，其尝试获取本地访问。R2L攻击通常于U2R攻击结合在一起。R2L攻击的一个例子是SSHbruteforce。

根据上面对异常分类以及网络攻击分类的介绍，按照文献[2]的划分，如图2．1所示，可以将四类网络攻击对应至三类异常。拒绝服务攻击由于是一系列大量的连接的同时活动，因此属于集合异常。探测属于在某些目的的驱使下收集特定信息，因此属于上下文异常。一般将U2R与R2L划分为点异常。

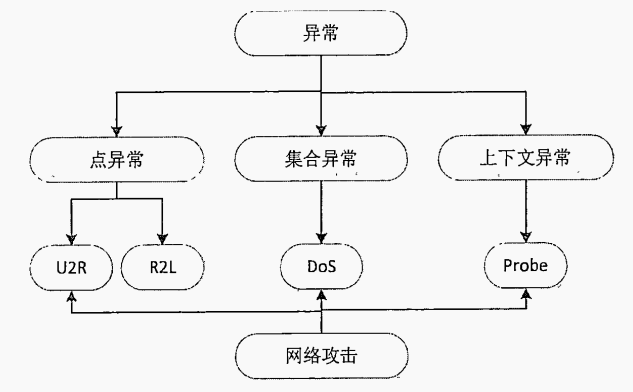
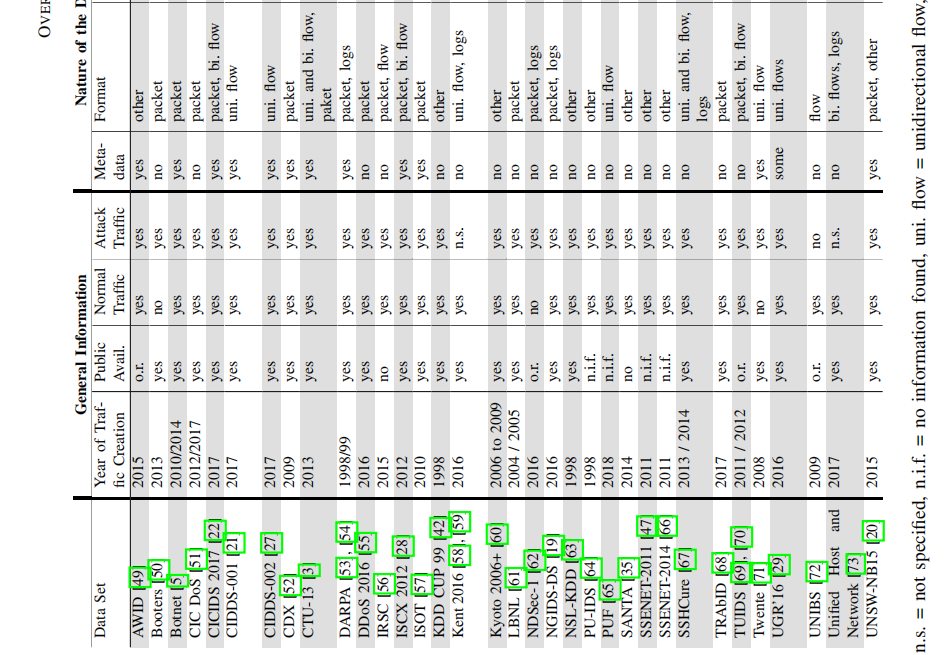
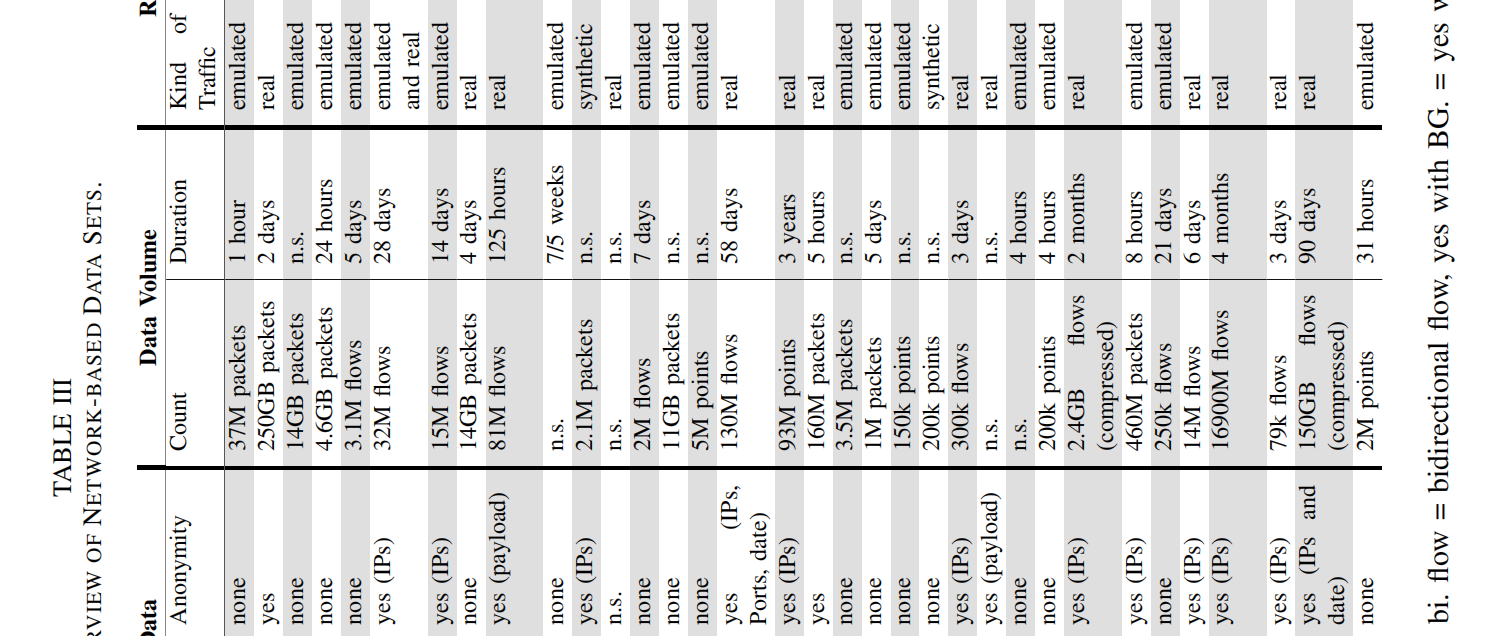
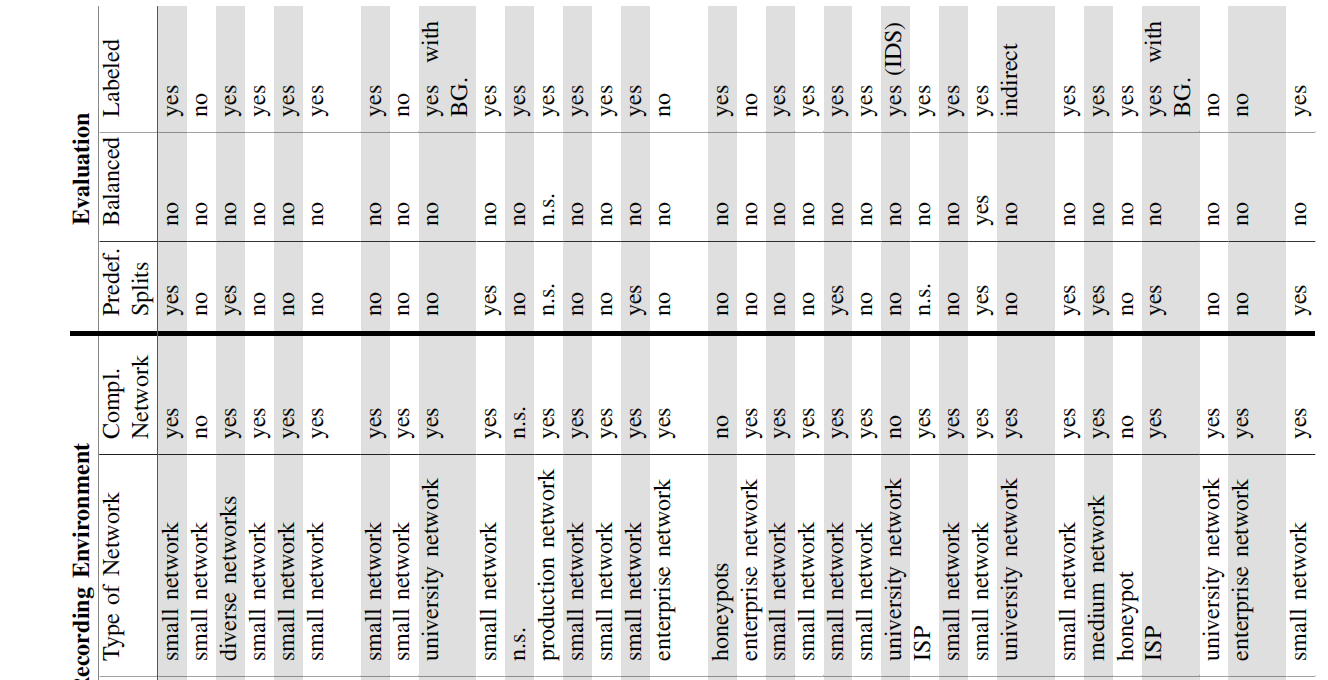


图1异常与网络攻击形式的对应关系图

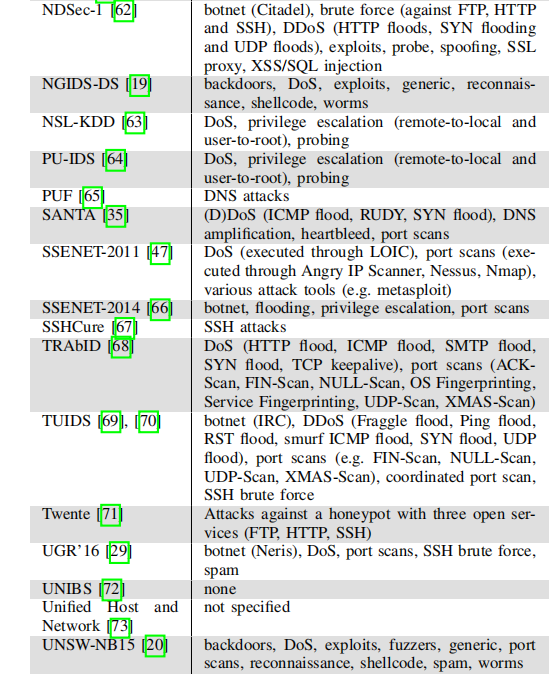
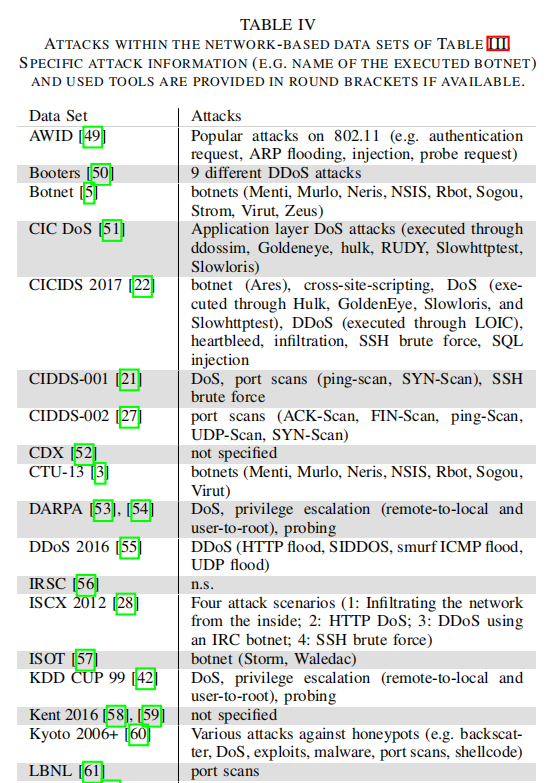
网络流量异常检测，是指应用各种异常检测技术分析网络流量，及时发现网络攻击行为。需要指出的是，和网络流量异常检测非常相近的一个概念是入侵检测，它们之间的区别如下。入侵检测，分为基于主机的入侵检测和基于网络的入侵检测，前者是指通过监视主机日志信息来检测入侵行为，例如对文件系统，注册表的改动信息等，后者是指通过分析进出的网络流量判定入侵行为。基于网络的入侵检测可分为误用检测和异常检测，误用检测是指基于已有攻击提取的流量指纹库进行匹配检测。异常检测是指根据历史数据建立机器学习模型以判定新流量数据是为异常，基于异常的网络入侵检测基本等同于安全相关的网络流量异常检测。

1. 异常检测数据集基本情况



表一 各个数据集基本情况

上表是关于现有数据库的基本情况。Genernal information（年份/是否公开/是否有正常数据/是否有异常数据）；Nather of the data（是否有元数据/数据类型/匿名性）；Date volume（数量/时长）；Recording environment（流量类型是否真实/网络类型/是否是完整网络）；Evaluation（是否有测试集，训练集/是否每个类别数据均衡/是否带标签）。



表二 各个数据集包含的攻击类型

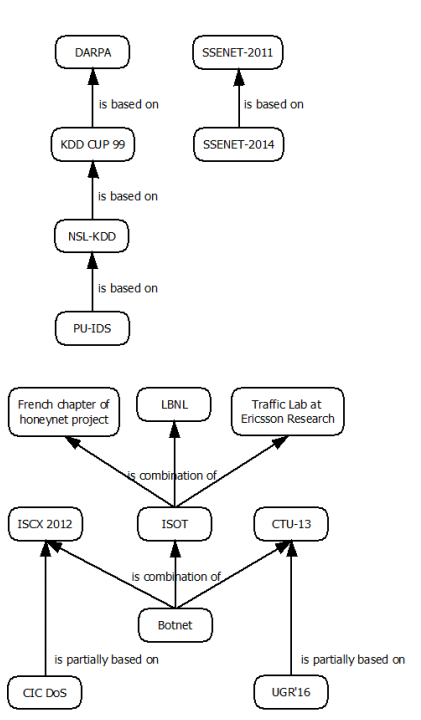


图2 各个数据集间的关系

三、各个数据集链接

AWID [icsdweb.aegean.gr/awid/index.html](https://links.jianshu.com/go?to=http://icsdweb.aegean.gr/awid/index.html" \t "_blank)

Booters [www.simpleweb.org/wiki/index.php](https://links.jianshu.com/go?to=https://www.simpleweb.org/wiki/index.php" \t "_blank)

Botnet [www.unb.ca/cic/datasets/botnet.html](https://links.jianshu.com/go?to=http://www.unb.ca/cic/datasets/botnet.html" \t "_blank)

CIC DoS [www.unb.ca/cic/datasets/dos-dataset.html](https://links.jianshu.com/go?to=http://www.unb.ca/cic/datasets/dos-dataset.html" \t "_blank)

CICIDS-2017 [www.unb.ca/cic/datasets/ids-2017.html](https://links.jianshu.com/go?to=http://www.unb.ca/cic/datasets/ids-2017.html" \t "_blank)

CIDDS-001 [www.hs-coburg.de/cidds](https://links.jianshu.com/go?to=http://www.hs-coburg.de/cidds" \t "_blank)

CIDDS-002 [www.hs-coburg.de/cidds](https://links.jianshu.com/go?to=http://www.hs-coburg.de/cidds" \t "_blank)

CDX [www.usma.edu/crc/sitepages/datasets.aspx](https://links.jianshu.com/go?to=https://www.usma.edu/crc/sitepages/datasets.aspx" \t "_blank)

CTU-13 [mcfp.weebly.com/the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html](https://links.jianshu.com/go?to=https://mcfp.weebly.com/the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html" \t "_blank)

DARPA [www.ll.mit.edu/ideval/docs/index.html](https://links.jianshu.com/go?to=https://www.ll.mit.edu/ideval/docs/index.html" \t "_blank)

DDoS 2016 [www.researchgate.net/publication/292967044\_Dataset-\_Detecting\_Distributed\_Denial\_of\_Service\_Attacks\_Using\_Data\_Mining\_Techniques](https://links.jianshu.com/go?to=https://www.researchgate.net/publication/292967044_Dataset-_Detecting_Distributed_Denial_of_Service_Attacks_Using_Data_Mining_Techniques" \t "_blank)

IRSC  not publicly available

ISCX 2012 [www.unb.ca/cic/datasets/ids.html](https://links.jianshu.com/go?to=http://www.unb.ca/cic/datasets/ids.html" \t "_blank)

ISOT [www.uvic.ca/engineering/ece/isot/datasets/](https://links.jianshu.com/go?to=https://www.uvic.ca/engineering/ece/isot/datasets/" \t "_blank)

KDD CUP 99 [kdd.ics.uci.edu/databases/kddcup99/kddcup99.html](https://links.jianshu.com/go?to=http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html" \t "_blank)

Kent 2016 [csr.lanl.gov/data/cyber1/](https://links.jianshu.com/go?to=https://csr.lanl.gov/data/cyber1/" \t "_blank)

Kyoto 2006+ [csr.lanl.gov/data/cyber1/](https://links.jianshu.com/go?to=https://csr.lanl.gov/data/cyber1/" \t "_blank)

LBNL [icir.org/enterprise-tracing/](https://links.jianshu.com/go?to=http://icir.org/enterprise-tracing/" \t "_blank)

NDSec-1 [www2.hs-fulda.de/NDSec/NDSec-1/](https://links.jianshu.com/go?to=http://www2.hs-fulda.de/NDSec/NDSec-1/" \t "_blank)

NGIDS-DS [research.unsw.edu.au/people/professor-jiankun-hu](https://links.jianshu.com/go?to=https://research.unsw.edu.au/people/professor-jiankun-hu" \t "_blank)

NSL-KDD [www.unb.ca/cic/datasets/nsl.html](https://links.jianshu.com/go?to=http://www.unb.ca/cic/datasets/nsl.html" \t "_blank)

PU-IDS  no link found

PUF  no link found

SANTA  not publicly available

SSENet-2011  no link found

SSENet 2014  no link found

SSHCure [www.simpleweb.org/wiki/index.php](https://links.jianshu.com/go?to=https://www.simpleweb.org/wiki/index.php" \t "_blank)

TRAbID [secplab.ppgia.pucpr.br/trabid](https://links.jianshu.com/go?to=https://secplab.ppgia.pucpr.br/trabid" \t "_blank)

TUIDS [agnigarh.tezu.ernet.in/~dkb/resources.html](https://links.jianshu.com/go?to=http://agnigarh.tezu.ernet.in/~dkb/resources.html" \t "_blank)

Twente [www.simpleweb.org/wiki/index.php](https://links.jianshu.com/go?to=https://www.simpleweb.org/wiki/index.php" \t "_blank)

UGR'16 [nesg.ugr.es/nesg-ugr16/index.php](https://links.jianshu.com/go?to=https://nesg.ugr.es/nesg-ugr16/index.php" \t "_blank)

UNIBS 2009 [netweb.ing.unibs.it/~ntw/tools/traces/](https://links.jianshu.com/go?to=http://netweb.ing.unibs.it/~ntw/tools/traces/" \t "_blank)

Unified Host and Network Data Set [csr.lanl.gov/data/2017.html](https://links.jianshu.com/go?to=https://csr.lanl.gov/data/2017.html" \t "_blank)

UNSW-NB15 [www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/](https://links.jianshu.com/go?to=https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/" \t "_blank)

四、对比较常用的几个数据集属性进行介绍

1. 以DARPA数据为基础的KDD CUP 99以及KLS-KDD(应用广，年代久，攻击全，但一些攻击已经过时)

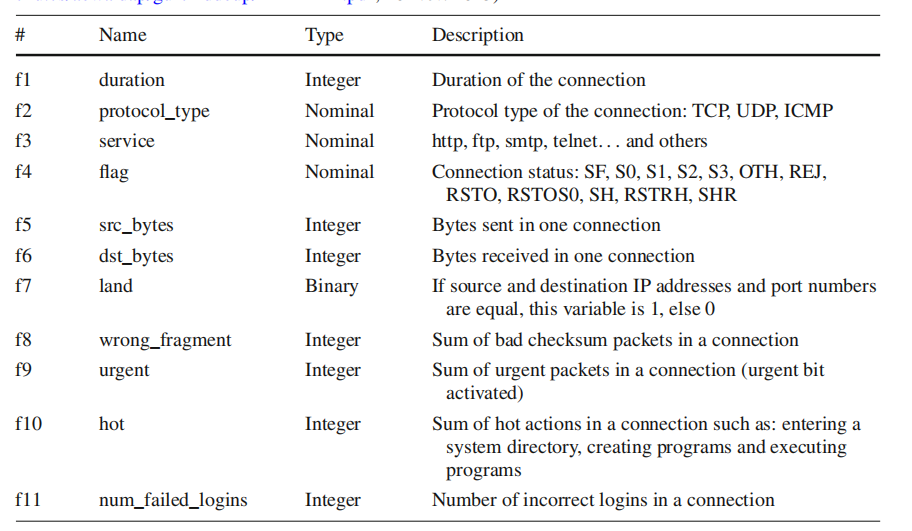
DARPA [www.ll.mit.edu/ideval/docs/index.html](https://links.jianshu.com/go?to=https://www.ll.mit.edu/ideval/docs/index.html" \t "_blank)（源数据）

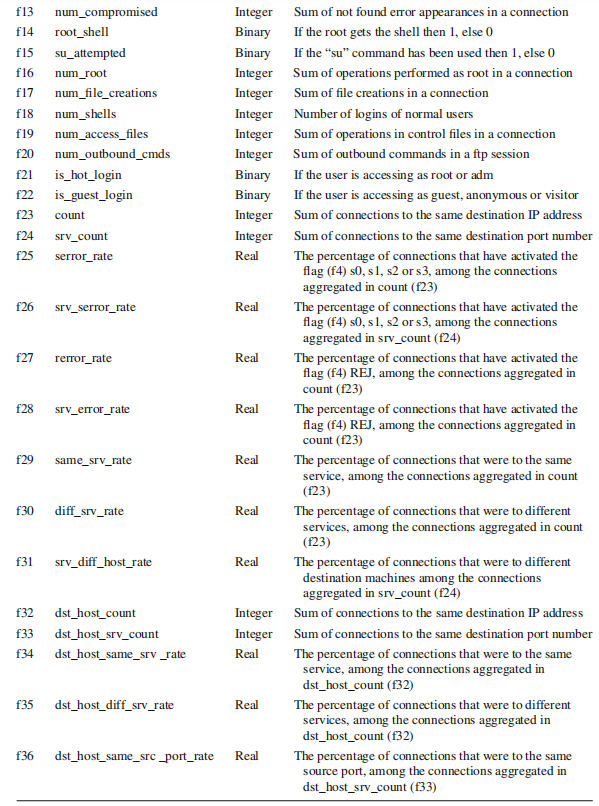
KDD CUP 99 [kdd.ics.uci.edu/databases/kddcup99/kddcup99.html](https://links.jianshu.com/go?to=http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html" \t "_blank)

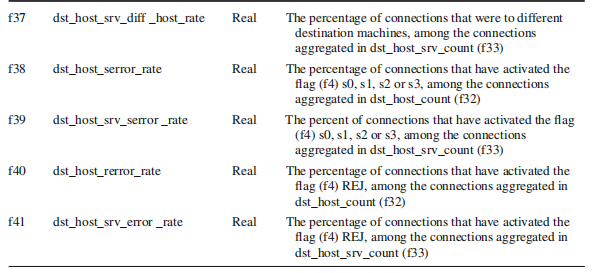
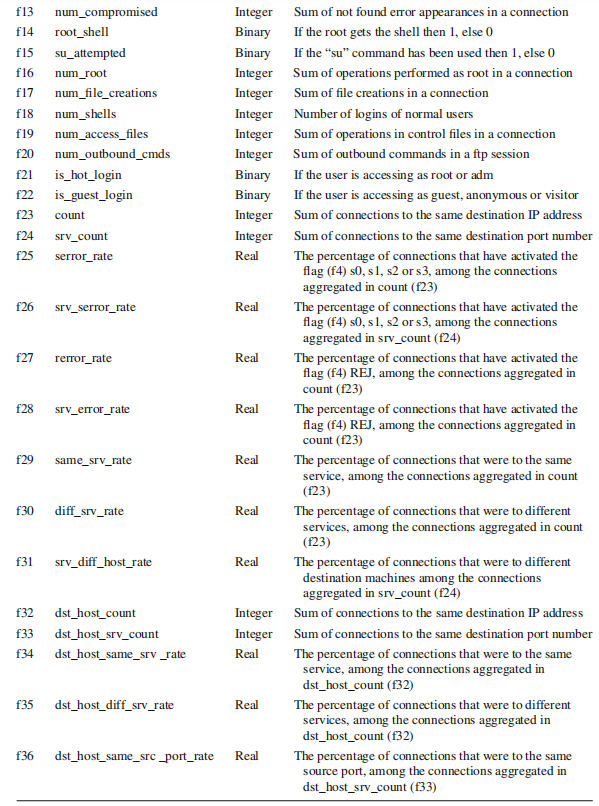
NSL-KDD [www.unb.ca/cic/datasets/nsl.html](https://links.jianshu.com/go?to=http://www.unb.ca/cic/datasets/nsl.html" \t "_blank)

实验室：美国国防部高级规划署于林肯实验室进行的一项美国空军局域网入侵检测评估项目

一共41个属性，下表为数据集





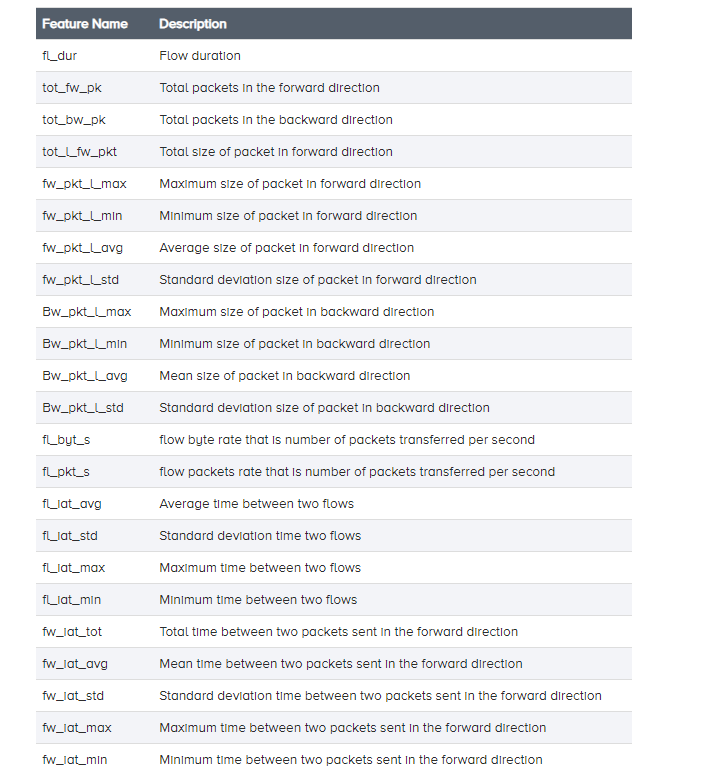


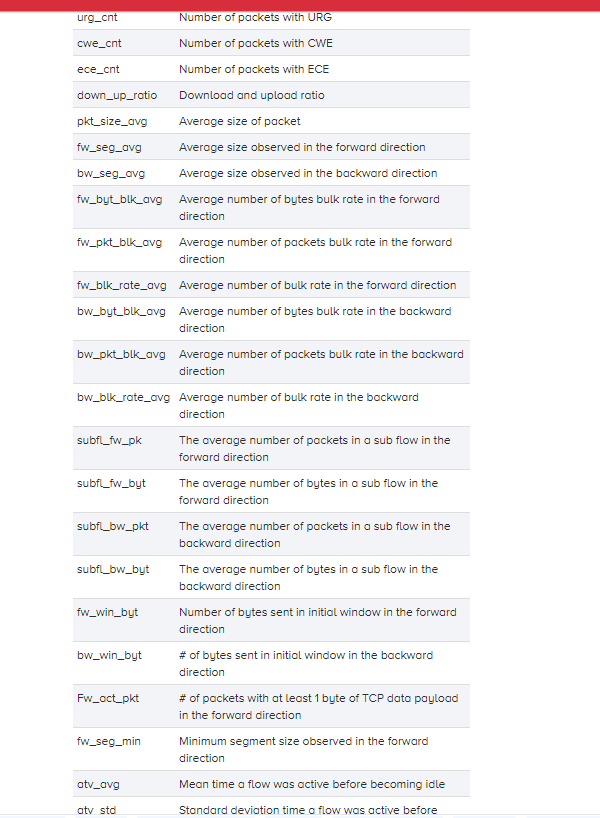
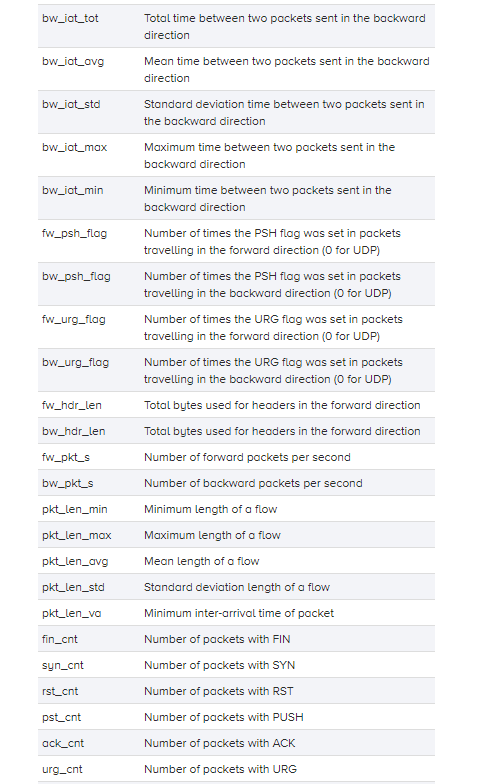
1. CICIDS-2017

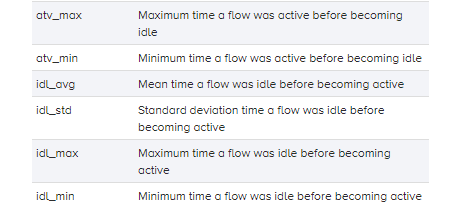
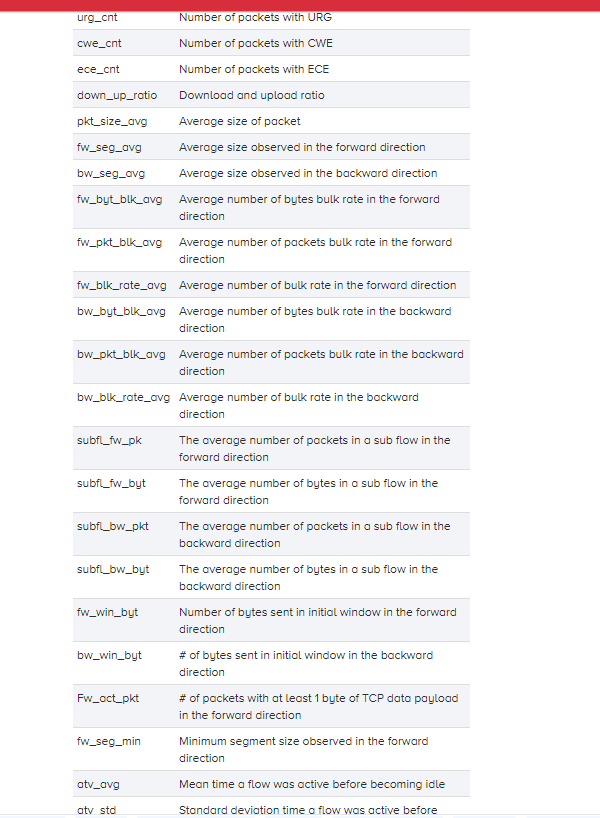
<https://www.unb.ca/cic/datasets/ids-2017.html>

实验室：加拿大网络安全研究所（CIC）（有pcap源文件）

提取的属性（有一个工具可以用来提取以下属性）

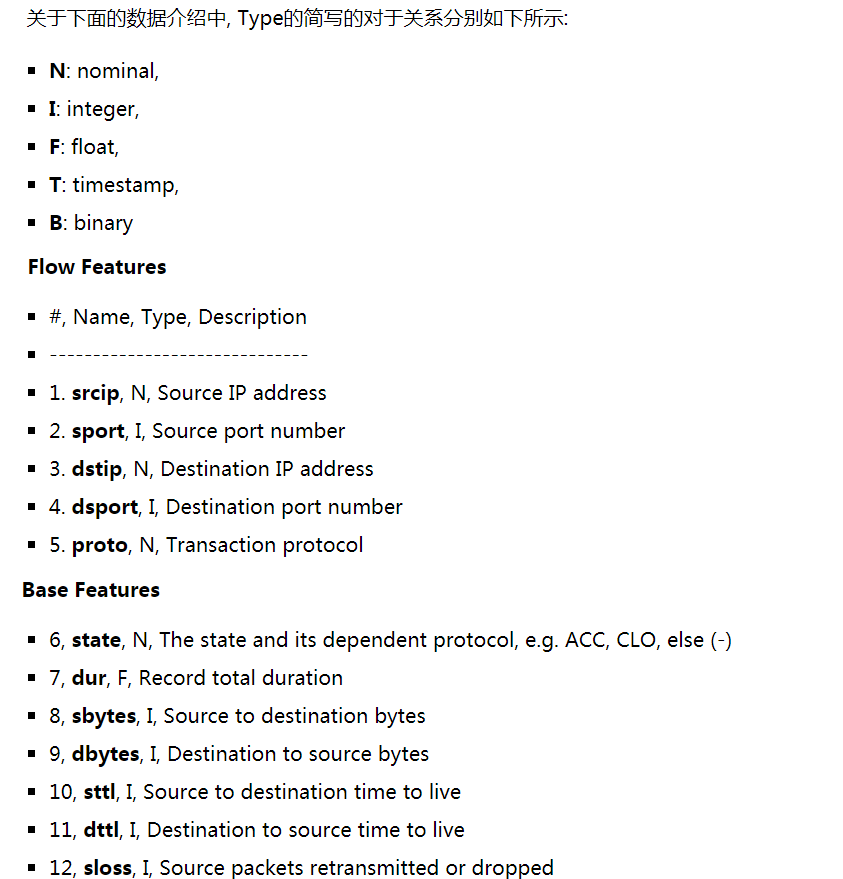


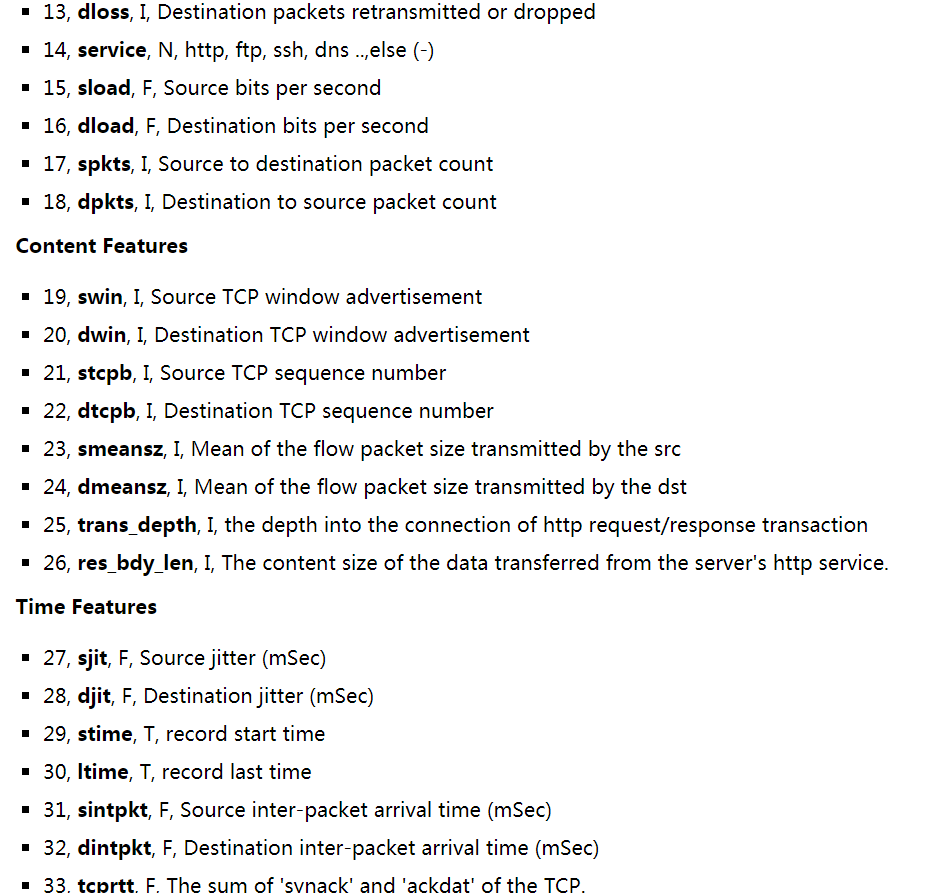


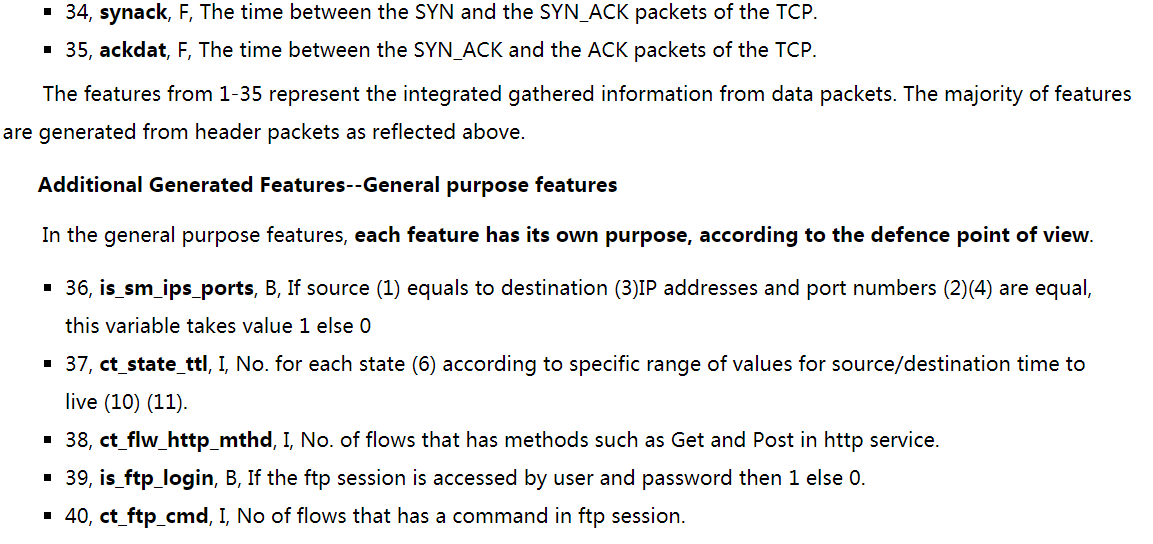


1. UNSW-NB15 [www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/](http://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/)

实验室：澳大利亚网络安全中心(ACCS)网络范围实验室







1. 一些感受和理解
2. 从数据包提取属性其实不是一件很容易的事情，如果对数据包的负载内容进行特征提取很复杂。
3. 进行实验时应该根据目的对数据包的特征进行筛选。

参考文献

1. Iglesias F , Zseby T . Analysis of network traffic features for anomaly detection[J]. Machine Learning, 2015, 101(1-3):59-84.
2. A Survey of Network-based Intrusion Detection Datata Sets