清华大学学位论文 LaTeX 模板
使用示例文档 v7.1.0

(申请清华大学工学硕士学位论文)

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 薛 瑞 尼

指 导 教 师 ： 郑 纬 民　教 授

副指导教师 ： 陈 文 光　教 授

二〇二一年三月

An Introduction to LaTeX Thesis Template of Tsinghua University v7.1.0

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Master of Science

in

Computer Science and Technology

by

Xue Ruini

Thesis Supervisor: Professor Zheng Weimin

Associate Supervisor: Professor Chen Wenguang

March, 2021

学位论文指导小组、公开评阅人和答辩委员会名单

指导小组名单

李 XX	教授	清华大学
王 XX	副教授	清华大学
张 XX	助理教授	清华大学

公开评阅人名单

刘 XX	教授	清华大学
陈 XX	副教授	XXXX 大学
杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所

答辩委员会名单

主席	赵 XX	教授	清华大学
委员	刘 XX	教授	清华大学
	杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所
	黄 XX	教授	XXXX 大学
	周 XX	副教授	XXXX 大学
秘书	吴 XX	助理研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）按照上级教育主管部门督导、抽查等要求，报送相应的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘　要

论文的摘要是对论文研究内容和成果的高度概括。摘要应对论文所研究的问题及其研究目的进行描述，对研究方法和过程进行简单介绍，对研究成果和所得结论进行概括。摘要应具有独立性和自明性，其内容应包含与论文全文同等量的主要信息。使读者即使不阅读全文，通过摘要就能了解论文的总体内容和主要成果。

论文摘要的书写应力求精确、简明。切忌写成对论文书写内容进行提要的形式，尤其要避免“第 1 章……；第 2 章……；……”这种或类似的陈述方式。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语。关键词不超过 5 个，每个关键词中间用分号分隔。

关键词：关键词 1；关键词 2；关键词 3；关键词 4；关键词 5

Abstract

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words “the first chapter”, “the second chapter” and the like should be avoided in the abstract.

Keywords are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 keywords, with semi-colons used in between to separate one another.

Keywords: keyword 1; keyword 2; keyword 3; keyword 4; keyword 5

目 录

插图和附表清单

符号和缩略语说明

PI	聚酰亚胺
MPI	聚酰亚胺模型化合物，N-苯基邻苯酰亚胺
PBI	聚苯并咪唑
MPBI	聚苯并咪唑模型化合物，N-苯基苯并咪唑
PY	聚吡咙
PMDA-BDA	均苯四酸二酐与联苯四胺合成的聚吡咙薄膜

第 1 章　引言

1.1　研究背景

网络自诞生以来，随着数十年的发展，已经成为了最重要的信息化基础设施之一。如图 1.1所示，第 46 次《中国互联网络发展状况统计报告》[1] 指出，截至 2020 年 6 月，我国网民规模达到 9.40 亿，互联网普及率达 67.0%，互联网应用涵盖即时通讯、搜索引擎、网络新闻、在线教育、购物出行等方面，可以说互联网已经和每个人的生活息息相关密不可分。

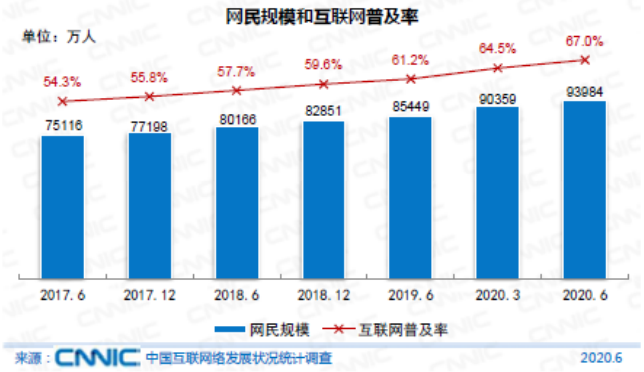

图 1.1　网民规模和互联网普及率

随着网络重要性的提升、用户规模的膨胀，管理网络的难度也越来越大。网络是一个复杂的系统，它的部署、运行和维护都需要专业的运维人员。早期的运维工作大部分是由运维人员手工完成，然后人们逐渐发现一些重复性的工作可以用自动化脚本来实现，于是诞生了自动化运维。自动化运维可以认为是基于专家经验、人为制定规则的系统。但是随着互联网规模急剧膨胀，以及服务类型的多样化，简单的、基于人为制定规则的方法并不能解决大规模运维的问题，因此产生了智能运维。与自动化运维依赖专家知识、人工生成规则不同，智能运维强调使用机器学习算法从海量运维数据中不断学习、不断提炼规则。

异常检测是智能运维的关键环节，具有至关重要的意义。从网络故障管理的角度来说，做好异常检测可以提前预测故障的发生；从性能管理的角度来说，可以发现性能不佳的区域，避免因误配置、架构不合理导致性能下降；从安全管理的角度来说，在网络攻击的前期阶段，及时发现并预警后续攻击，进而做出防御措施。因此，在复杂的网络环境中甄别出有效和异常流量尤为重要，在重大事故发生前，根据各项流量特征的变化，提前预测出即将发生的事故，提高应急响应速度，防患于未然。

1.2　研究内容

1.3　主要贡献

1.4　论文组织结构

第 2 章　相关工作综述

2.1　引言

本章对网络流量异常检测领域的相关工作进行综述。首先介绍校园网网络流量的特点，

异常检测是一个重要的领域，自 1980 年以来，国内外已经有无数学者在这方面做研究。分类、统计、信息理论和聚类。

Ahmed et al.[2] 将异常检测技术分为分类、统计、信息理论和聚类四类。

本章还讨论了用于网络入侵检测的数据集的研究挑战。

2.2　网络流量异常的定义和分类

Hawkins(1980) 给出了异常的本质性的定义[3]：异常是在数据集中与众不同的数据，使人怀疑这些数据并非随机偏差，而是产生于完全不同的机制。例如在道路交通领域，某条道路的车流量突然增多甚至堵塞或者突然减少，此时车流量数据就是一个异常。因此网络行为的异常就是指那些偏离了正常的、标准的或预期的行为。为了检测网络异常，网络所有者必须有一个预期或正常行为的概念，我们称其为基线。要检测网络行为的异常，就需要持续监控网络中的意外趋势或事件，那些可能改变网络流量特征或者监控指标的恶意行为。

因此我们只关注引起网络流量特征变化的恶意行为，而对于系统权限提升，缓冲区溢出等黑客攻击手段不做研究。网络异常根据

网络流量异常具体有哪些类别，学术界没有统一的意见。本文中关注的网络流量异常按照产生意图分为恶意和非恶意两类，其中恶意行为主要有拒绝服务攻击，网络扫描，BGP 劫持，网络蠕虫，僵尸网络等；非恶意行为主要有物理故障，突发事件等。接下来我们对这些异常分别进行介绍：

1. 拒绝服务攻击：拒绝服务 (DoS) 攻击是一种网络攻击，恶意行为者的目的是通过中断计算机或其他设备的正常运作，使其目标用户无法使用该设备。DoS 攻击的功能通常是通过构造大量请求淹没目标机器，直到正常的流量无法处理，导致额外用户的拒绝服务。分布式拒绝服务（DDoS）攻击是一种来自许多分布式来源的 DoS 攻击，如僵尸网络 DDoS 攻击。因其攻击成本低、攻击效果明显等特点，DDoS 攻击仍然是互联网用户面临的最常见、影响较大的网络安全威胁之一。DoS 攻击通常分为 2 类。（1）缓冲区溢出攻击：一

种攻击类型，内存缓冲区溢出会导致机器消耗所有可用的硬盘空间、内存或 CPU 时间。这种形式的利用通常会导致行为迟缓、系统崩溃或其他有害的服务器行为，导致拒绝服务。（2）洪范攻击：通过用大量的数据包使目标服务器饱和，恶意行为者能够使服务器容量过饱和，导致拒绝服务。为了使大多数 DoS 泛滥攻击成功，恶意行为者必须拥有比目标更多的可用带宽。

2. 网络扫描。网络扫描黑客在进行网络攻击之前，首先要进行网络扫描，从中寻找可以攻击的目标。具体来说，黑客需要确定网络中哪些主机是活动的，活动主机上运行了哪些存在漏洞的服务等，从而决定下一步的攻击计划。网络扫描分为主机扫描和端口扫描两种。

   主机扫描主机扫描的目的是确定网络中存在哪些在线的主机或设备。端口扫描端口扫描的目的是确定活动主机上开放了哪些端口，运行了哪些网络服务。按照扫描方式来分，端口扫描分为垂直扫描和水平扫描。垂直扫描会扫描某个主机的所有端口，而水平扫描则扫描网络中所有主机的某个固定端口。
3. BGP 劫持。BGP 劫持是指攻击者恶意地对互联网流量进行重新路由。攻击者通过谎称拥有一组 IP 地址（称为 IP 前缀）的所有权来实现这一目的，而实际上他们并不拥有、控制或路由。BGP 劫持就像有人把高速公路上的所有标志都换掉，把汽车流量改道到错误的出口。
4. 网络蠕虫。网络蠕虫不同于计算机病毒。与计算机病毒不同的是，计算机蠕虫不需要附在别的程序内，可能不用用户介入操作也能自我复制或运行。计算机蠕虫未必会直接破坏被感染的系统，却几乎都对网络有害。计算机蠕虫可能会执行垃圾代码以发动分布式拒绝服务攻击，令计算机的执行效率极大程度降低，从而影响计算机的正常使用；可能会损毁或修改目标计算机的文件；亦可能只是浪费带宽。（恶意的）计算机蠕虫可根据其目的分成 2 类：

   一种是面对大规模计算机使用网络发动拒绝服务的计算机蠕虫，虽说会绑架计算机，但用户可能还可以正常使用，只是会被占用一部分运算、连网能力。另一种是针对个人用户的以执行大量垃圾代码的计算机蠕虫。计算机蠕虫多不具有跨平台性，但是在其他平台下，可能会出现其平台特有的非跨平台性的平台版本。第一个被广泛注意的计算机蠕虫名为：“莫里斯蠕虫”，由罗伯特 · 泰潘 · 莫里斯编写，于 1988 年 11 月 2 日释出第一个版本。这个计算机蠕虫间接和直接地造成了近 1 亿美元的损失。这个计算机蠕虫释出之后，引起了各界对计算机蠕虫的广泛关注。
5. 僵尸网络。僵尸网络 (简称” 机器人网络”) 是指由感染了恶意软件的计算机组成的网络，这些计算机由单一攻击方控制，即所谓的” 僵尸继承者”。在”

机器人携带者” 控制下的每一台机器都被称为” 机器人”。攻击方可以从一个中心点，指挥其僵尸网络上的每一台计算机同时进行协调的犯罪行动。僵尸网络的规模 (许多僵尸网络由数百万个僵尸组成) 使攻击者能够进行大规模的行动，这在以前的恶意软件中是不可能的。由于僵尸网络一直处于远程攻击者的控制之下，受感染的机器可以接收更新，并在飞行中改变其行为。因此，僵尸牧民往往能够在黑市上租用其僵尸网络部分的访问权，以获取大量经济利益。

6. 物理故障是指路由器故障，链路破坏，断电等不可预测的突发事件。该类异常以两种方式影响着网络链路的流量。
7. 突发事件。突发事件往往是正常的网络操作。

2.3　网络流量异常检测算法

本节将介绍在异常检测领域主流的一些算法，根据所依赖的技术原理的不同，将这些算法分为了基于统计、基于分类、基于聚类、基于信息论、基于深度学习的异常检测算法。

2.3.1　基于分类的异常检测算法

基于分类的技术依赖于专家对网络攻击特征的广泛了解。当网络专家向检测系统提供详细的特征时，具有已知模式的攻击一经发起就能被检测出来。这完全依赖于攻击的签名，作为一个系统，只有当网络专家较早地提供了攻击的签名，它才能够检测出攻击。这说明一个只能够检测到它所知道的系统很容易受到新的攻击，而新的攻击会不断出现不同的版本，并且更加隐蔽地发起。即使创建了新的攻击的签名并将其纳入系统中，最初的损失也是不可替代的，而且修复程序非常昂贵。

基于分类的方法依赖于建立知识库的正常流量活动特征，并将偏离基线特征的活动视为异常活动。其优势在于它们能够检测到完全新颖的攻击，假设它们表现出大量的偏离正常配置文件的情况。此外，由于知识库中未包含的正常流量被认为是攻击，因此会有无意中的误报。因此，异常检测技术需要进行训练，以建立正常的活动配置文件，这很耗时，而且还取决于是否有完全正常的流量数据集。在实践中，获得无攻击的流量实例是非常罕见且昂贵的。此外，在当今 s 动态和不断变化的网络环境中，保持正常配置文件的更新是非常困难的。在现有的大量基于分类的网络异常检测技术中，我们主要讨论以下四种技术。

Eskin et al.[4] 引入无监督 SVM 的概念来检测异常事件。常规的 SVM 的原理

是推导出一个超平面，使得正类样本和负类样本之间的分离余量最大化，将特征空间中的两类数据进行分离。标准的 SVM 算法是一种监督学习方法，需要标记数据来创建分类规则。而该算法经过改进 SVM，试图将整个训练数据集从原点分离出来，找到一个以最大余量将数据实例与原点分离的超平面，

Hu et al.[5] 提出了一种忽略噪声数据的异常检测方法，该方法使用 Robust SVM(RSVM) 来开发。标准的 SVM 有一个主要假设，即所有的训练样本数据都是独立且相同分布的 (i.i.d)。但是在实际场景中，训练数据往往包含噪声，这就会导致标准 SVM 会学习出一个高度非线性的决策边界，从而导致通用性较差。基于此，RSVM 以类中心的形式加入了平均化技术，使得决策面更加平滑。此外，RSVM 另一个优点是能大大降低支持向量的数量，从而减少运行时间，提高效率。

Kruegel et al.[6] 假设异常检测系统包含许多模型，用于分析一个事件的不同特征，指出了在这种系统下异常检测技术造成高误报率的两个主要原因。一是异常检测系统通过将多个概率模型的输出进行汇总，而每个模型往往只给出一个事件的常态/异常的得分或概率，从而导致高误报率；二是异常检测系统无法处理那些不正常但合法的行为，如 CPU 利用率、内存使用率突然增高等。基于贝叶斯网络的概念，Kruegel et al.[6] 提出了一种解决上述问题的方法。对于一个输入事件的有序流 ($\boldsymbol{S} = e_1, e_2, e_3...$)，异常检测系统决策每个事件是正常还是异常。该决策基于 k 个模型 ($\boldsymbol{M} = m_1, m_2, ..., m_k$) 的输出 ($o_i | i = 1, 2, ..., k$) 和可能的附加信息 ($\boldsymbol{I}$)。应用贝叶斯网络来识别异常事件，引入根节点，根节点代表一个具有两种状态的变量。一个子节点用于捕捉模型 的输出，子节点与根节点相连，预计当输入异常或正常时，输出事件会有所不同。贝叶斯网络最近的应用可以在电信网络中找到（Deljac 等，2015）。贝叶斯网络是对包含不确定性的领域进行建模的一种有效方法。一个离散的随机变量用一个有向无环图（DAG）表示，其中每个节点反映了随机变量的状态，并包含一个条件概率表（CPT）。CPT 的任务是提供一个节点处于特定状态的概率。在贝叶斯网络中，节点之间存在着父子关系，这表明子节点所代表的变量依赖于父节点所代表的变量。由于这种网络可以用于事件分类方案，因此也适用于网络异常检测。

神经网络已经被应用于各个应用领域，如图像和语音处理，但其对计算量的要求很高。神经网络对数据进行分类的优势也可被用于网络异常检测。在网络异常检测领域，神经网络通常会和其他技术进行结合，如统计方法。2020outlier 提出了一个多层的前馈神经网络，该神经网络可以用来进行异常值的检测。具体来说，Replicator Neural Networks 是一个多层前馈的神经网络 (multi-layer feed-forward neural networks)，在输入层和输出层之间放置了三个隐藏层。它的目标是通过训练

在输出层以最小的误差重现输入数据模式。由于该模型中间隐藏层节点的个数少于输入输出层节点的个数，这样就起到了压缩数据和恢复数据的作用。RNN 的示意图如图 2.1所示。

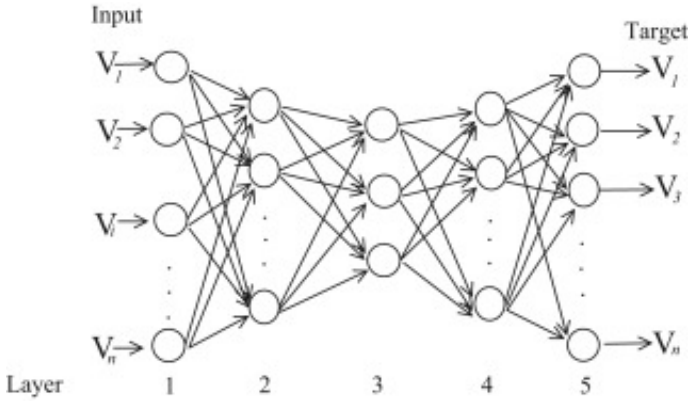

图 2.1　Replicator Neural Networks 示意图

Zhang[7] 提出了一种神经网络与统计模型相结合的分层入侵检测系统。将神经网络分类器的输出表示为一个连续变量 (t)，其中-1 表示有绝对把握的入侵，1 表示没有攻击。此外，自组织地图 (SOM) 被用于网络异常检测。Ramadas 等 (2003) 提出，利用 SOM，可以对网络流量进行实时分类。SOM 依赖于这样一个假设，即网络攻击可以由不同的神经元组来描述，这些神经元组与其他神经元组相比，在输出神经元图上覆盖更大的区域。Poojitha et al.[8] 开发了一个由反向传播算法训练的前馈神经网络，利用给定的数据集与计算机网络在正常和异常行为期间的相关信息来检测异常。

2.3.2　基于统计的异常检测算法

早期的异常检测方法往往基于统计与概率模型，也就是假设-检验的方法。首先对数据的分布做出假设，然后找出假设下所定义的“异常”，因此往往使用极值分析或假设检验。比如对最简单的一维数据假设服从正态分布，然后将距离均值某个范围以外的点当做异常点。推广到高维后，假设各个维度相互独立。这类方法的好处速度一般比较快，但是因为存在很强的“假设”，效果不一定很好。

入侵检测技术也是利用统计理论发展起来的；例如，Ye 和 Chen(2001 年) 在异常检测中使用了成熟的 chi-square 理论。根据这种技术，建立了一个信息系统中正常事件的档案。这种方法的基本思想是既要检测出与正常事件有较大偏差的异常事件，又要检测出入侵事件。基于 chi-square 检验统计量的距离测量方法被开发为

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i} \tag{2-1}$$

其中 X_i 为第 i 个变量的观测值，E_i 为第 i 个变量的期望值，n 为变量的数量。

当一个变量的观测值接近预期时，χ^2 的值就会很低。根据 3σ 定律，当观测值的 χ^2 大于 $\bar{X^2} + 3S_X^2$ 时，该值被视为异常。

Krügel et al.[9] 提出了一种用于检测异常网络流量的统计处理单元，更具体地说，是为了检测 R2L 和 U2R 等罕见的攻击。开发了一种度量方法，使系统能够自动搜索不同服务请求的相同特征。根据以下三个主要特征计算出请求的异常得分。请求的类型；请求的长度；以及有效载荷分布。网络管理员定义了一个阈值，以便对异常请求发出警报。异常得分的计算方法如式 (5)，其中有效负载分布的权重大于其他属性。(5) 基于统计学理论的原理，我们开发了不同类型的技术来检测异常，接下来将讨论。

在时间序列异常检测领域，最常见的基于统计的算法为 ARIMA，即差分自回归移动平均模型 [8]。我们将流量信号分解为两部分，一是遵循一定规律、可预测的正常变化，二是由突发性变化组成、不可预测的异常情况。ARIMA 分析和建模用于网络流量预测，能够检测和识别流量异常或异常值。

2.3.3　基于信息论的异常检测算法

信息理论的测量方法可以用来建立一个适当的异常检测模型。在 Lee 和 Xiang(2001) 的一篇论文中，使用了几种度量方法，如熵、条件熵、相对熵、信息增益和信息成本来解释数据集的特征。我们对这些度量方法的定义如下。

熵是信息论的一个基本概念，它衡量一个数据项集合的不确定性。对于一个数据集 D，其中每个数据项都属于一个类 (x)，D 相对于分类的熵定义为：1. (13) 其中 P(x) 为 x 在 D 中的概率。

条件熵是指 D 的熵，给定 Y 是概率分布 () 的熵，为 (14) 其中 P(x，y) 是 x 和 y 的联合概率和 x 给定 y 的条件概率。

相对熵是指定义在相同的两个概率分布 p(x) 和 q(x) 之间的熵，具体为 (15)

相对条件熵是指定义在同一和上的两个概率分布（和）之间的熵。(16)

信息增益是对数据集 D 中某一属性或特征 A 的信息增益的衡量，是 (17) 其中，值 A 为 A 的可能值集，Dv 为 D 的子集，其中 A 的值为 v。

Ambusaidi 等人（2014）中提出了一种基于非线性相关系数（NCC）的相似性测量方法，以提取网络流量之间的线性和非线性相关性。提取的相关信息用于检测恶意网络行为。Pearson s 相关系数是一种基本的线性相关方法，用于找出两个变量之间的依赖关系 (Ahmed 等，2015c)，然而，在一些数据集中，不同变量之间存在非线性相关，如网络流量中。NCC 由 Wang 等 (2005) 定义，如式 (18)，其中和为变量 X 和 Y 的修正熵。

给定一组 m 个正常训练数据实例，首先计算 NCC。对于任何传入实例，传入实例与正常实例之间的 NCC 记录为。对于用户定义的阈值 σ，其范围在 0 和 1 之间，如果 NCC 的差异大于 σ，则认为一个传入流量实例是异常的（19）。

在 Tan 等 (2014a) 中，针对 DoS 攻击检测，提出了一个利用多元相关分析 (MCA) 的系统，通过提取网络流量特征之间的几何相关性，来实现网络流量的精确特征分析。检测过程主要包含三个步骤，如图 6 所示。在步骤 1 中，在一个明确的时间区间内生成基本特征。第 2 步包含多元相关分析，应用” 三角区域图生成” 模块，提取第一步得出的每个流量实例中两个不同特征之间的相关性。第三步是基于训练和测试阶段的决策。

基于这些知识，可以建立适当的异常检测模型。有监督的异常检测技术需要先有一个训练数据集，再有一个测试数据来评估模型的性能。在这种情况下，首先，使用信息理论措施来确定模型是否适合测试新数据集。Noble 和 Cook(2003) 在基准 DARPA 和 UNM 审计数据集上进行了实验，以证明信息理论措施的效用，并得出结论，它们可以用来创建高效的异常检测模型，也可以用来解释它们的性能

Tan 等人 (2014a) 中的多变量相关分析方法的概念被纳入到网络流量实例的表征中，并将其转换为相应的图像。这些图像被用于 DoS 攻击检测，基于一个广泛使用的异构度量，即地球移动者距离（Earth Mover s Distance，EMD）（Rubner 等，1998）。EMD 考虑了跨区域匹配，比其他一些著名的异同度测量方法更准确地评估了分布之间的异同度。

2.3.4　基于聚类的异常检测算法

聚类指的是无监督学习算法，它不需要预先标记数据来提取相似数据实例的分组规则（Jain 等，1999）。虽然有不同类型的聚类技术，但我们讨论常规聚类和共聚类对网络异常检测的有用性。常规聚类和共聚类的区别在于行和列的处理。常规聚类技术如 k-means（Ahmed 和 Naser，2013）考虑数据集的行进行聚类，而共聚类则同时考虑数据集的行和列来产生聚类（Ahmed 等人，2015d）。

下面简单讨论一下使用聚类检测异常时总是要做的三个关键假设。假设 1：由于我们只能创建正常数据的聚类，因此，后续任何与现有正常数据聚类不相适应的新数据都被认为是异常数据；例如，由于基于密度的聚类算法不包括聚类内的噪声（Ester 等人，1996），噪声被认为是异常数据。

假设 2：当一个簇同时包含正常数据和异常数据时，已经发现正常数据靠近最近的簇中心点，但异常数据远离中心点（Ahmed 和 Naser，2013）。在这种假设下，异常事件使用距离得分来检测。

假设 3：在一个具有不同大小的聚类中，较小和较稀疏的聚类可以被认为是异

常的，较厚的聚类是正常的。属于大小和/或密度低于阈值的聚类的实例被认为是异常的。

Münz 等人 (2007) 对异常数据采用的方法非常直接。他们使用 k-means 聚类来生成正常和异常聚类。一旦实现聚类，就使用以下假设进行分析。

如果一个实例比异常簇中心点更接近正常，则该实例被列为正常，反之亦然。

如果实例与中心点之间的距离大于预定义的阈值 (dmax)，则该实例被视为异常；以及

如果一个实例比正常聚类中心点更接近异常聚类中心点，或者它与正常聚类中心点的距离大于预定义的阈值，则被视为异常。

Petrovic 等 (2006) 提出了一种基于聚类评价技术组合的聚类标签策略。将 Davies-Bouldin 聚类评价指数和聚类中心直径的比较结合起来，以充分应对攻击向量的特性。他们考虑了相应聚类的紧凑性和它们之间的分离度，以及区分分析网络中" 正常" 和" 异常" 行为的主要参数。然而，他们并没有解释他们的 k-means 聚类使用 k=2 的原因。根据他们的方法，攻击向量通常非常相似，如果不是完全相同的话；例如，在大规模攻击的情况下，相应的聚类是非常紧凑的，这种聚类的 Davies-Bouldin 指数要么是 0（当非攻击聚类是空的时候），要么是非常接近 0。考虑到攻击向量之间的预期相似性，因为攻击聚类的中心点的直径预期比非攻击聚类的直径小，他们可以区分正常和异常的聚类。

Portnoy 等 (2001) 提出了基于宽度的聚类来对数据实例进行分类。宽度是恒定的，对所有聚类都保持不变。一旦进行聚类，基于正常实例在整个数据集中占压倒性比例的假设，N% 的聚类是正常的，其余是异常的。利用这一假设，Leung 和 Leckie（2005）提出了一种基于密度和网格的聚类算法，该算法适用于无监督的异常检测。

Syarif 等人（2012）研究了各种聚类算法应用于异常检测时的性能。他们使用了五种不同的方法，即 k-means、改进的 k-means、k-medoids、期望值最大化（EM）聚类和基于距离的异常检测算法。表 3 展示了用于网络异常检测的聚类算法的性能评价。

聚类算法通常是基于距离/密度发现异常点。基于距离/密度的异常点检测方法的关键步骤在于给每个数据点都分配一个离散度，其主要思想是：针对给定的数据集，对其中的任意一个数据点，如果在其局部邻域内的点都很密集，那么认为此数据点为正常数据点，而异常点则是距离正常数据点最近邻的点都比较远的数据点。通常有阈值进行界定距离的远近。异常检测领域下常用的聚类算法有 k-means、LOF、孤立森林、高斯混合模型 [20] 等。

2.3.5　基于深度学习的异常检测算法

随着深度学习的兴起，越来越多的学者尝试用深度学习算法来进行异常检测，尤其是针对时间序列数据，深度学习模型往往表现出惊人的效果。常用的深度学习算法为变分编码器、神经网络 [6][14]、生成对抗网络、LSTM[17]、RNN[3][4][10][12][13][15] 等。以变分自动编码器 (Variational Auto-Encoder)[5] 为例，其利用自编码器的重构误差和局部误差，针对时间序列的异常检测的场景，达到了很好的效果。

2.4　异常检测领域开源数据集介绍

数据集主要由 KDDCUP99, CICIDS 等。网络流量异常检测领域最为经典的数据集当属 KDD99，但是这个数据集年代过于久远，对于现在的网络环境早已不适用。NSL-KDD 是为了解决 KDD'99 数据集的一些固有问题而提出的数据集，这些问题在 [1] 中已经提到。虽然，这个新版本的 KDD 数据集仍然存在 McHugh 所讨论的一些问题，并且可能不能完美地代表现有的真实网络，但由于缺乏基于网络的 IDS 的公共数据集，我们相信它仍然可以作为一个有效的基准数据集来帮助研究人员比较不同的入侵检测方法。

此外，NSL-KDD 训练集和测试集的记录数量是合理的。这一优势使得在完整的集合上运行实验是经济实惠的，而不需要随机选择一小部分。因此，不同研究工作的评价结果将具有一致性和可比性。

CICIDS2017 数据集包含了良性的和最新的常见攻击，与真实的现实世界数据（PCAPs）相似。它还包括使用 CICFlowMeter 进行网络流量分析的结果，并根据时间戳、源和目的 IP、源和目的端口、协议和攻击（CSV 文件）对流量进行了标注。同时还提供了提取的特征定义。

生成真实的背景流量是我们构建这个数据集的首要任务。我们使用了我们提出的 B-Profile 系统（Sharafaldin，等人，2016）来对人类交互的抽象行为进行剖析，并生成自然的良性背景流量。对于这个数据集，我们基于 HTTP、HTTPS、FTP、SSH 和电子邮件协议建立了 25 个用户的抽象行为。

数据采集期从 2017 年 7 月 3 日（周一）上午 9 点开始，到 2017 年 7 月 7 日（周五）下午 5 点结束，共 5 天。其中周一为正常日，只包括良性流量。实施的攻击包括蛮力 FTP、蛮力 SSH、DoS、Heartbleed、Web 攻击、渗透、僵尸网络和 DDoS。它们在周二、周三、周四和周五的上午和下午都被执行过。

在我们最近的数据集评估框架中 (Gharib 等人，2016)，我们确定了建立一个可靠的基准数据集所必需的 11 个标准。之前的 IDS 数据集都无法覆盖这 11 项标

准的全部内容。在下文中，我们简要地概述了这些标准。

THU-IDS 清华校园网数据集，该数据集为真实流量，将于第三章进行介绍。

2.5　异常检测算法对比

对比不同机器学习方法在 NSL-KDD 数据集上的效果，

2.6　现有异常检测算法存在的问题

第 3 章　基于深度学习的时间序列异常检测算法研究

3.1　引言

3.2　循环神经网络原理

第 4 章　数学符号和公式

4.1　数学符号

研究生《写作指南》要求量及其单位所使用的符号应符合国家标准《国际单位制及其应用》(GB 3100—1993)、《有关量、单位和符号的一般原则》(GB/T 3101—1993) 的规定。模板中使用 unicode-math 宏包来配置数学符号，与 LaTeX 默认的英美国家的符号习惯有所差异：

1. 大写希腊字母默认为斜体，如 `\Delta`：Δ。
2. 有限增量符号 Δ（U+2206）应使用 unicode-math 宏包提供的 `\increment` 命令。
3. 向量、矩阵和张量要求粗斜体，应该使用 unicode-math 的 `\symbf` 命令，如 `\symbf{A}`、`\symbf{\alpha}`。
4. 数学常数和特殊函数要求用正体，应使用 `\symup` 命令，如 $\symup{\pi} = 3.14\cdots$；$\mathrm{e} = 2.718\cdots$，
5. 微分号和积分号使用使用正体，比如 $\int f(x)\,\mathrm{d}x$。

关于数学符号更多的用法，参考 unicode-math 宏包的使用说明，全部数学符号命的令参考 unimath-symbols。

关于量和单位推荐使用 siunitx 宏包，可以方便地处理希腊字母以及数字与单位之间的空白，比如：$6.4\times10^{6}\,\mathrm{m}$，$9\,\mu\mathrm{m}$，$\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-1}$，$10\,^{\circ}\mathrm{C}\sim20\,^{\circ}\mathrm{C}$。

4.2　数学公式

数学公式可以使用 equation 和 equation* 环境。注意数学公式的引用应前后带括号，建议使用 `\eqref` 命令，比如式 (4-1)。

$$\frac{1}{2\pi\mathrm{i}}\int_{\gamma} f = \sum_{k=1}^{m} n(\gamma;a_k)\mathscr{R}(f;a_k) \tag{4-1}$$

注意公式编号的引用应含有圆括号，可以使用 `\eqref` 命令。

多行公式尽可能在“=”处对齐，推荐使用 align 环境。

$$a = b + c + d + e \tag{4-2}$$

$$= f + g \tag{4-3}$$

4.3　数学定理

定理环境的格式可以使用 amsthm 或者 ntheorem 宏包配置。用户在导言区载入这两者之一后，模板会自动配置 thoerem、proof 等环境。

定理 4.1 (Lindeberg–Lévy 中心极限定理)：　设随机变量 $X_1, X_2, \cdots, X_n$ 独立同分布，且具有期望 μ 和有限的方差 $\sigma^2 \neq 0$，记 $\bar{X}_n = \frac{1}{n}\sum_{i+1}^{n} X_i$，则

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \leqslant z\right) = \Phi(z), \tag{4-4}$$

其中 $\Phi(z)$ 是标准正态分布的分布函数。

证明：　Trivial. ∎

同时模板还提供了 assumption、definition、proposition、lemma、theorem、axiom、corollary、exercise、example、remar、problem、conjecture 这些相关的环境。

第 5 章　引用文献的标注

模板支持 BibTeX 和 BibLaTeX 两种方式处理参考文献。下文主要介绍 BibTeX 配合 natbib 宏包的主要使用方法。

5.1　顺序编码制

在顺序编码制下，默认的 `\cite` 命令同 `\citep` 一样，序号置于方括号中，引文页码会放在括号外。统一处引用的连续序号会自动用短横线连接。

参考文献

[1] 中国互联网信息中心. 中国互联网络发展状况统计报告 [Z].

[2] Ahmed M, Mahmood A N, Hu J. A survey of network anomaly detection techniques[J]. Journal of Network and Computer Applications, 2016, 60: 19-31.

[3] Hawkins D M. Identification of outliers: volume 11[M]. Springer, 1980.

[4] Eskin E, Arnold A, Prerau M, et al. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data[M]. 2002: 77-101.

[5] Hu W, Liao Y, Vemuri R. Robust support vector machines for anomaly detection in computer security.[C]// 2003: 168-174.

[6] Kruegel C, Mutz D, Robertson W, et al. Bayesian event classification for intrusion detection[C]// 19th Annual Computer Security Applications Conference, 2003. Proceedings. IEEE, 2003: 14-23.

[7] Zhang Z. Hide : a hierarchical network intrusion detection system using statistical preprocessing and neural network classification[C]// Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY, 5-6 June, 2001. 2001.

[8] Poojitha G, Kumar K N, Reddy P J. Intrusion detection using artificial neural network[C/OL]// 2010 Second International conference on Computing, Communication and Networking Technologies. 2010: 1-7. DOI: 10.1109/ICCCNT.2010.5592568.

[9] Krügel C, Toth T, Kirda E. Service specific anomaly detection for network intrusion detection[C]// the 2002 ACM symposium. 2002.

附录 A　补充内容

附录是与论文内容密切相关、但编入正文又影响整篇论文编排的条理和逻辑性的资料，例如某些重要的数据表格、计算程序、统计表等，是论文主体的补充内容，可根据需要设置。

A.1　图表示例

A.1.1　图

附录中的图片示例（图 A.1）。

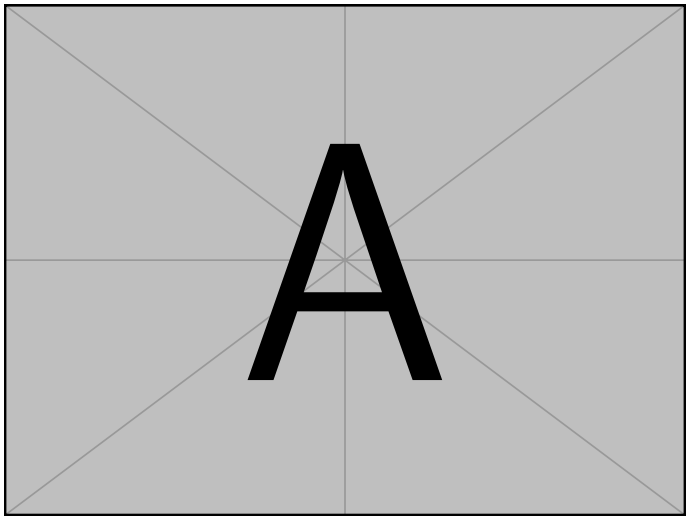

图 A.1　附录中的图片示例

A.1.2　表格

附录中的表格示例（表 A.1）。

A.2　数学公式

附录中的数学公式示例（公式 (A-1)）。

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathscr{R}(f; a_k) \tag{A-1}$$

表 A.1　附录中的表格示例

文件名	描述
thuthesis.dtx	模板的源文件，包括文档和注释
thuthesis.cls	模板文件
thuthesis-*.bst	BibTeX 参考文献表样式文件
thuthesis-*.bbx	BibLaTeX 参考文献表样式文件
thuthesis-*.cbx	BibLaTeX 引用样式文件

致　谢

衷心感谢导师 ××× 教授和物理系 ×× 副教授对本人的精心指导。他们的言传身教将使我终生受益。

在美国麻省理工学院化学系进行九个月的合作研究期间，承蒙 Robert Field 教授热心指导与帮助，不胜感激。

感谢 ××××× 实验室主任 ××× 教授，以及实验室全体老师和同窗们学的热情帮助和支持！

本课题承蒙国家自然科学基金资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____