第5章 基于大数据分析的入侵检测技术

5.1 入侵检测技术

早期计算机面对网络攻击一般只有用防火墙抵御攻击。随着网络安全技术的发展，诞生了入侵检测技术。入侵检测技术是对防火墙技术发展的合理补充，它充分扩展了在网络空间领域内对于人员和系统的全面控制能力。近年来，全球范围内影响网络空间安全的黑客入侵事件数量急剧增加，网络空间变得越来越不安全，使得入侵检测技术日益凸显其重要性。历经30余年的发展，入侵检测技术已经从最初的一种有价值的研究想法和单纯的理论模型发展成为具有种类繁多的实际原型和软硬件产品的实用网络空间安全技术。

本章首先简要介绍入侵检测技术的发展、未来展望、系统组成和入侵检测的方法，然后介绍大数据分析技术在入侵检测中的应用。

5.1.1 入侵检测技术概述

入侵就是指可能会对计算机系统或者计算机网络系统的安全造成威胁的一系列连续的恶意行为，包括对未经授权的信息访问、对重要信息数据的篡改以及拒绝服务攻击等。入侵检测是指对入侵行为进行诊断、识别并做出响应的过程。实施入侵检测的系统称为入侵检测系统(Intrusion Detection System，IDS)。

1. 入侵检测系统的发展

入侵检测技术历经了由理论概念提出到进行技术研究再到实用产品开发的发展历程，包含概念诞生、模型发展、百家争鸣、继续演进4个具体的发展阶段。

1) 概念诞生阶段

1980年，James P.Anderson为一个保密客户写了一份名为《计算机安全威胁监控与监视》的技术报告，第一次详细阐述了入侵检测的概念。该报告提出了一种针对计算机系统所面临的风险和威胁进行分类的方法，并将威胁分为外部渗透、内部渗透和不法行为3种，还提出了利用审计跟踪数据监视入侵活动的思想，这份报告被公认为入侵检测概念的诞生标志。

入侵检测的研究历史最早可追溯到 20 世纪 80 年代，但是因为当时互联网还处于起步阶段，使用者数量较少，因此入侵检测概念的诞生并没有引起很大的关注。随着互联网技术飞速发展，黑客攻击事件频发，个人隐私数据价值提升，用户数据增长，品牌的价值观念受到重视，使得人们对于入侵检测技术的重视有很大提高，入侵检测技术也得到了迅速发展。

2）模型发展阶段

1984—1986 年，乔治敦大学的 Denning 和 SRI 公司计算机科学实验室的 Neumann 设计了一个实时入侵检测系统模型，将其命名为入侵检测专家系统（Intrusion Detection Expert System，IDES）。该系统模型独立于特定的系统平台、应用环境、系统弱点以及入侵类型，为构建入侵检测系统提供了一个通用框架。它是第一个在应用软件中运用了统计和基于规则两种技术的系统模型。1987 年，Denning 在前期工作的基础之上，发表论文《一种入侵检测模型》，文中提出了入侵检测的基本模型，并介绍了几种可用于入侵检测的统计分析模型。该论文被认为是入侵检测领域研究工作的开端，该论文中提出的模型是入侵检测研究领域的里程碑。此后大量的入侵检测系统模型开始出现，很多都是基于 Denning 的统计量分析理论。

3）百家争鸣阶段

1990 年，加州大学戴维斯分校的 Heberlein 发表论文《一种网络安全监控器》，该论文设计的网络安全监控器第一次将网络数据包作为入侵检测的信息源，它利用捕获的 TCP/IP 信息将数据分组，监控异构网络环境中的异常活动。该系统将网络数据流作为入侵检测分析的数据来源，可以在不将审计数据转换成统一格式的情况下检测异常主机，具有很好的通用性。该成果第一次构建了基于网络的入侵检测技术，为入侵检测技术的发展史开辟了新的篇章。此外，随着 Porras 和 Kemmerer 基于状态转换分析的入侵检测技术的提出和完善，根据已知攻击模型进行入侵检测的方法成为该领域研究的另一热点。

4）继续演进阶段

1990 年以后，集成了实用入侵检测技术的商用入侵检测产品逐步出现，并形成了更多的入侵检测技术。早期的入侵检测技术仅执行网络攻击监测或者提供有限的数据分析功能。随着技术的发展，出现了许多新型的入侵检测技术，有的增加了应用层数据分析的能力，有的能够配合防火墙进行联动，从而和防火墙功能互补，有效地阻断攻击事件，形成了入侵防御的功能。

2. 入侵检测技术的未来展望

入侵检测技术的未来会朝以下几个方向发展：

（1）大规模分布式入侵检测系统以及异构系统之间的协作和数据共享。结合分布式技术和网络技术，基于分布式的多层次入侵检测系统可以很好地解决这个问题，分布式网络环境下的入侵检测将成为未来研究的热点。

（2）入侵检测系统的自身保护。目前入侵检测面临自身安全性的挑战，一旦系统中的入侵检测部分被入侵者控制，整个系统的安全防线将面临崩溃的危险。对于如何防止入侵者对入侵检测系统功能的削弱乃至破坏的研究将在很长时间内持续下去。

(3) 入侵检测智能化。信息安全受到前所未有的挑战,单一的安全技术很难保证系统的真正安全。与神经网络、机器学习、深度学习等人工智能算法结合,使得入侵检测方法能够通过学习训练进而识别出新型未知攻击,同时降低误报率,将成为入侵检测技术未来的发展趋势之一。

(4) 建立入侵检测系统评价体系。设计通用的入侵检测测试、评估方法和平台,实现对多种入侵检测系统的检测,已成为当前入侵检测系统的另一重要研究与发展领域。评价入侵检测系统可从检测范围、系统资源占用、自身的可靠性等方面进行,评价指标有:能否保证自身的安全、运行与维护系统的开销、报警准确率、负载能力以及可支持的网络类型、支持的入侵特征数、是否支持 IP 碎片重组、是否支持 TCP 流重组等。

3. 入侵检测系统的组成

从系统组成上看,入侵检测系统一般包括 3 个模块:数据采集模块、入侵检测模块、响应模块。

数据采集模块根据入侵检测类型的不同采集不同类型的数据,例如网络的数据包、操作系统的系统调用日志或者应用日志。数据采集模块往往伴随着对采集到的信息进行预处理的过程,包括对信息进行简单的过滤,输出格式化的信息。前者有助于消除冗余数据,提升系统的性能;后者可提升系统的互操作性。经过预处理后的信息一般先存放到日志数据库中,再提交给入侵检测模块进行下一步的操作。

入侵检测模块由分析引擎和检测策略两部分组成。分析引擎内部具备检测算法功能,率先接收数据采集模块发送过来的数据。检测算法内部包含了从最简单的字符串匹配算法到复杂的专家系统甚至神经网络等一系列算法,用于对数据进行分析。因此,分析引擎是入侵检测系统的核心,分析引擎将会最终判定一个行为是异常的还是正常的。检测策略包含了如何诊断入侵的配置信息、入侵的签名(也就是入侵的行为特征)、各种阈值。入侵检测模块综合分析引擎和检测策略的判断,在对行为进行入侵判断后将判断的结果直接发给响应模块。

响应模块根据响应策略中预先定义的规则进行对应的响应处理。一般来说,可能的响应行为不仅包括发出报警通知管理员,而且可以对恶意行为自动地采取反击措施。这样,一方面可自动地重新配置目标系统,阻断入侵者;另一方面可以重新配置检测策略和数据采集策略,例如,可针对正在进行的特定行为采集更多的信息,对行为的性质进行更准确的判断。

5.1.2 入侵检测的方法

对于目前已有入侵检测的分类有多种方法,比较通用一种分类方法是基于检测原则的不同,将入侵检测系统划分为异常检测型入侵检测系统和误用检测型入侵检测系统。另一种入侵检测的分类方式是根据检测输入数据来源的不同,将入侵检测系统分为基于主机的入侵检测系统和基于网络的入侵检测系统。

1. 异常检测型入侵检测系统

异常检测也称为基于行为的入侵检测,以系统、网络、用户或进程的正常行为建立轮

廓模型(即正常行为模式),将与正常行为偏离较大的行为解释成入侵。在异常检测方法中,通常建立一个关于系统正常活动的状态模型,在进行入侵检测时,将用户当前的活动情况与这个正常模型进行对比,如果检测到两者的差异超过可以容忍的程度,则将这个行为标识为入侵行为。采用异常检测方法的入侵检测技术不需要对每种入侵行为进行定义,就可以用于检测未知的入侵行为,不需要维护庞大的网络攻击特征库。但是,利用该方法检测到新型的入侵行为时,可能无法准确判定该入侵行为是何种类型。相对于误用检测方法,该方法检测结果的漏报率低,但误报率高,其难点在于如何不把正常的活动误认为入侵以及忽略真正的入侵行为。

异常检测方法具有检测未知攻击的能力,但是由于新攻击方法总是不断出现,因此异常检测技术一直较受重视,产生了大量的异常检测技术。下面对其中的主要技术进行介绍和分析。

1) 基于统计分析的检测方法

统计分析是异常检测的主要方法之一。该方法依据系统中特征变量的历史数据建立统计模型,并运用该模型对特征变量未来的取值进行预测和偏离检测。系统中的特征变量有用户登录失败次数、CPU 和 I/O 利用率、文件访问数及访问出错率、网络连接数、击键频率、事件的时间间隔等。

统计分析方法包括以下模型:

(1) 均值与标准差模型是以单个特征变量为检测对象,假定特征变量满足正态分布,根据该特征变量的历史数据统计出分布参数(均值、标准差),并依此设定信任区间。在检测过程中,若特征变量的取值超出信任区间,则认为发生异常。

(2) 多元模型是以多个特征变量为检测对象,分析多个特征变量间的相关性,是均值与标准差模型的扩展,不仅能检测到单个特征变量值的偏离,还能检测到特征变量间关系的偏离。

(3) 马尔可夫过程模型将每种类型的事件定义为系统的一个状态,用状态转换矩阵表示状态的变化,若对应于发生的事件的状态转移概率较小,则该事件可能为异常事件。

(4) 时间序列模型将事件计数与资源消耗根据时间排成序列,如果某一新事件在相应时间发生的概率较低,则该事件可能为入侵。

以统计分析方法形成系统或用户的行为轮廓实现简单,且在度量选择得较好(即系统或用户行为的变化会在相应的度量上产生显著的变化)时能够可靠地检测出入侵。该方法的缺点是:需要以系统或用户一段时间内的行为特征作为检测对象,因此检测的时效性差,一般在将特定的行为判定为入侵时,可能已经发生了入侵事件,已经造成了经济财产的损失;度量的阈值难以确定;忽略了事件间的时序关系。

2) 基于数据挖掘的检测方法

数据挖掘是一种利用分析工具在大量数据中提取隐含在其中且潜在有用的信息和知识的方法。入侵检测过程也是基于采集的大量数据信息,如主机系统日志、审计记录和网络数据包等,对数据信息进行分析,从而发现入侵或异常的过程。因此,可利用数据挖掘技术,从大量数据中提取尽可能多的、隐藏的安全信息,抽象出有利于比较和判断的特征模型(如基于异常检测的正常行为轮廓)。

数据挖掘方法有多种，运用到入侵检测中的主要有关联分析、序列分析和聚类分析3种。其中，关联分析方法主要分析事件记录中数据项间隐含的关联关系，形成关联规则；序列分析方法主要分析事件记录间的相关性，形成事件记录的序列模式；聚类分析方法识别事件记录的内在特性，将事件记录分组以构成相似类，并导出事件记录的分布规律。在建立上述关联规则、序列模式和相似类后，即可依此检测入侵或异常行为。

基于数据挖掘的检测方法建立在对采集的大量信息进行分析的基础之上，只能进行事后分析，即只有入侵事件发生后才能检测到入侵行为的存在。

3）其他检测方法

其他异常检测方法有基于规则的方法、人工免疫法、基于机器学习的检测方法和基于神经网络的检测方法等。

异常检测的优点为：不需获取攻击特征，能检测未知攻击或已知攻击的变种，且能适应用户或系统等行为的变化。但异常检测具有如下缺点：一般根据经验知识选取或不断调整阈值以满足系统要求，阈值难以设定；异常不一定由攻击引起，系统易将用户或系统的特殊行为（如出错处理等）判定为入侵，同时系统的检测准确性受阈值的影响，在阈值选取不合适时，会产生较多的检测错误，造成检测错误率高；攻击者可逐渐修改用户或系统行为的轮廓模型，因而检测系统易被攻击者训练；无法识别攻击的类型，因而难以采取适当的措施阻止攻击的继续。

2. 误用检测型入侵检测系统

误用检测也称为特征检测、基于知识或基于签名的入侵检测。它将已知的入侵活动用一种模式表示，形成网络攻击特征库，或称为网络攻击规则库。该方法对输入的待分析数据源进行适当处理，提取其特征，并将这些特征与网络攻击特征库中的特征进行比较，如果发现匹配的特征，则指示发生了一次入侵行为。根据入侵检测数据来源的不同，网络攻击特征的含义也随之不同。采用误用检测方法的入侵检测技术的误报率低、漏报率高，能够准确地识别已知的攻击，并可以详细地报告入侵攻击类型。但是，该方法对新的入侵方法无能为力，需要不断地将新的入侵模式加入到网络攻击特征库中，才能提高其识别新网络攻击的能力。

常用的误用检测技术主要有以下3种。

1）基于专家系统的检测方法

基于专家系统的检测方法是常用的入侵检测方法，通过将有关入侵的知识转化为IF-THEN结构的规则，IF…为构成入侵的条件，THEN…为发现入侵后采取的解决措施。基于专家系统的检测方法优点在于分离了系统的推理控制过程和问题的最终解答，即用户不需要理解或干预专家系统内部的推理过程，只需把专家系统看作一个黑盒。

在将专家系统应用于入侵检测时，存在下列问题：缺乏处理序列数据的能力，即无法处理数据的前后相关性；检出攻击的性能很大程度上取决于设计者的知识；只能检测已知的攻击模式；无法处理和判断不确定性；规则库难以维护，更改规则时要考虑对规则库中其他规则的影响。

2）基于状态转移分析的检测方法

基于状态转移分析的检测方法运用状态转换图表示和检测已知的攻击模式，即运用系统状态和状态转移表达式描述已知的攻击模式，以有限状态机模型表示入侵过程。入侵过程由一系列导致系统从初始状态转移到入侵状态的行为组成，其中，初始状态为入侵发生前的系统状态，入侵状态表示入侵完成后系统所处的状态。

用于误用检测的状态转移分析引擎包括一组状态转移图，各自代表一种入侵或渗透模式。每当有新行为发生时，状态转移分析引擎检查所有的状态转移图，查看是否会导致系统的状态转移。如果新行为否定了当前状态的断言，状态转移分析引擎将状态转移图回溯到断言仍然成立时的状态；如果新行为使系统状态转移到了入侵状态，状态转移信息就被发送到决策引擎，由决策引擎根据预定义的策略采取相应措施。

以状态转移分析方法表示的入侵检测过程只与系统状态的变化有关，而与入侵的过程无关。状态转移分析方法能检测到协同攻击和慢攻击；能在攻击行为尚未到达入侵状态时检测到该攻击行为，从而及时采取相应措施阻止攻击行为。状态转换图给出了保证攻击成功的特征行为的最小子集，能检测到具有相同入侵模式的不同表现形式。状态转移分析方法中状态对应的断言和特征行为需要手工编码，如果将其应用到复杂的入侵场景时会存在问题。

3）其他检测方法

其他的误用检测方法有基于有色 Petri 网的误用检测及基于键盘监控的误用检测等。

误用检测的优点为：攻击检测的准确率高；能够识别攻击的类型。误用检测的缺点为：只能检测已知攻击；滞后于新出现的攻击，对于新的攻击，仅在其包含到攻击特征库中后才能检测到；攻击特征库的维护较为困难，新攻击出现后需由专家根据专业知识抽取攻击特征，不断更新攻击特征库；攻击者可通过修改攻击行为，使其与攻击特征库中的特征不相符，从而绕过检测。

误用检测和异常检测各有优缺点，具有一定的互补性。很多研究试图将这两种技术结合起来，从而发挥两者的优点，提高检测效率，并且最大限度地减少错误率。

3. 基于主机的入侵检测系统

基于主机的入侵检测系统简称主机入侵检测系统，它通过对主机系统状态、事件日志和审计记录进行监控以发现入侵。基于主机的入侵检测系统保护的一般是它所在的系统。它利用的系统信息主要如下：

（1）用户行为，如登录时间、击键频率、击键错误率、输入命令等。

（2）系统状态，如 CPU 使用率、内存使用率、I/O 及硬盘使用率等。

（3）进程行为，如系统调用、CPU 使用、I/O 操作等。

（4）网络事件，在网络栈处理通信数据之后应用层程序处理之前对通信数据进行解释。

基于主机的入侵检测系统的优点为：能检测内部授权人员的误用以及成功避开传统的系统保护方法而渗透到网络内部的入侵活动；具有操作系统及运行环境的信息，检测准确性较高；在检测到入侵后可与操作系统协同阻止入侵的继续，响应及时。

基于主机的入侵检测系统的缺点为：与操作系统平台相关，可移植性差；需要在每个被检测主机上安装入侵检测系统，因此成本较高；难以检测针对网络的攻击，如消耗网络资源的 DoS 攻击、端口扫描等。

4. 基于网络的入侵检测系统

基于网络的入侵检测系统简称网络入侵检测系统，它监视网络段中的所有通信数据包，识别可疑的或包含攻击特征的活动。在广播网络中，可以将网卡设置为混杂模式，监控整个网络而不暴露自己的存在。基于网络的入侵检测系统能够利用网络数据中的许多特征，例如单个包的 TCP/IP 头、包的内容以及多个包的组合。

基于网络的入侵检测系统的优点为：对用户透明，隐蔽性好，使用简便，不容易遭受来自网络的攻击；与被检测的系统平台无关；仅需较少的探测器；往往用独立的计算机完成检测工作，不会给运行关键业务的主机带来额外负载；攻击者不易转移证据。

基于网络的入侵检测系统的缺点为：无法检测到网络内部的合法用户滥用系统；无法分析传输的加密报文；在交换式网络中不能保证实用性；易被攻击者绕过；系统对所有的网络报文进行分析，增加了相关主机的负担，且易受 DoS 攻击；入侵响应的延迟较大。

5.2 大数据分析技术在入侵检测中的应用

大数据分析技术一直是入侵检测技术中重要的数据分析方法，从入侵检测技术诞生之日起，大数据分析技术便已经开始在其研究和应用中发挥重要作用了。早在 1987 年，Denning 在提出入侵检测的基本模型时，就已经将大数据分析中的方法应用到入侵检测中了。

虽然入侵检测技术已经过多年的发展，但是仍然存在较严重的误报和漏报等不可忽视的问题，学术界和产业界也在寻求不同的技术方法提高入侵检测的质量。大数据是一个交叉研究领域，它采用统计学、数学、机器学习等方法，可以发现大型数据集中信息的未知模式和关系。近年来，大数据分析技术更加完善，这也恰好可以为降低传统入侵检测技术的误报和漏报问题提供解决途径，越来越多的先进技术将被应用到入侵检测的研究和实用系统中。在入侵检测技术中，常用的大数据分析算法有朴素贝叶斯算法、逻辑回归算法等。

5.2.1 朴素贝叶斯算法

1. 朴素贝叶斯算法概述

1）分类问题

人们对于分类问题其实并不陌生。例如，当看到一个陌生人时，会下意识判断他（她）是男是女，其实这就是一种分类操作。从数学角度来说，分类问题可做如下定义：已知集合 $C=\{y_1,y_2,\cdots,y_n\}$ 和 $I=\{x_1,x_2,\cdots,x_m\}$，确定映射规则 $y=f(x)$，使得对任意 $x_i\in I$ 有且仅有一个 $y_i\in C$ 使得 $y_i=f(x_i)$ 成立（不考虑模糊数学里的模糊集情况）。其中，C 称为类别集合，其中每一个元素是一个类别；而 I 称为项集合，其中每一个元素是一个待

分类项；f 称为分类器。分类算法的任务就是构造分类器 f。

这里要着重强调，对于分类问题，往往采用经验性方法构造映射规则，即一般情况下的分类问题缺少足够的信息构造 100％正确的映射规则，因而要通过对经验数据进行学习，从而实现一定概率意义上正确的分类，这样训练出的分类器并不是一定能将每个待分类项准确映射到相应的类上。分类器的质量与分类器构造方法、待分类数据的特性以及训练样本数量等诸多因素有关。例如，医生对病人进行诊断就是一个典型的分类过程，任何一个医生都无法直接看到病人的病情，只能通过观察病人表现出的症状和各种化验检测数据来推断病情，这时医生就好比一个分类器，而这个医生诊断的准确率与他当初受到的教育（构造方法）、病人的症状是否明显（待分类数据的特征）以及医生的经验多少（训练样本数量）都有密切关系。

2）贝叶斯分类算法简介

贝叶斯分类算法是一类算法的总称，这类算法均以贝叶斯定理为基础。在介绍贝叶斯分类算法前，首先介绍两个概念：经典概率和主观概率。经典概率是代表事件发生的客观概率，是客观存在的。而主观概率则是个人的主观估计，随个人的主观认识变化而变化。以抛硬币为例，经典概率表示抛硬币时硬币正面/反面朝上的概率，是客观存在的；而主观概率是个人相信硬币正面/反面朝上的概率。还有常见的一个例子，如果一台个人计算机频繁发生死机和重启的现象，那么一个安全工程师通常会认为这台计算机被感染病毒的概率是 90％，这里的 90％是安全工程师凭借自己的网络安全经验和计算机当前的现状综合判断得到的概率。贝叶斯概率就是主观概率，对它的估计取决于先验知识的正确以及后验知识的丰富和准确。贝叶斯概率常常随着个人掌握的信息不同而发生较大变化。例如，在一场足球赛中足球王国巴西队将迎战欧洲劲旅西班牙队，比赛前不同人对球赛胜负的预测都是不同的。假设通常人们会预测比赛胜负的比例为 1∶1（两队实力都很强）；如果知道巴西队的主力前锋球员在比赛前的训练中负伤，无缘比赛，那么比赛的胜负比例可能会调整为 1∶2；更进一步，如果知道西班牙队的主教练受到感冒影响无法率队，由助理教练代为指挥球队，那么比赛的胜负比例可能会调整为 2∶3，这里所有的预测都是主观的，是基于后验知识的一种判断，取决于对双方各种比赛信息的掌握。

3）相关基础知识

假设事件 A 发生的概率记为 $P(A)$，通常事件 A 在事件 B 发生的条件下的概率 $P(A|B)$ 与事件 B 在事件 A 发生的条件下的概率 $P(B|A)$ 是不一样的。

然而这两者没有确定的关系，贝叶斯定理就是对这种关系的陈述。贝叶斯定理认为，一个事件会不会发生取决于该事件在先验分布中已经发生过的次数。设 $P(A)$ 为已知事件 A 发生的概率，可以称之为先验概率（prior probability），例如假设手机中病毒的概率为 10％。而 $P(B|A)$ 是在考虑事件 A 之后对事件 B 发生的概率的估计，称为条件概率，条件概率是后验概率（posterior probability），例如手机中毒后导致手机中的金融账户产生财产损失的概率为 80％。这里条件概率的计算公式为

$$P(B \mid A)=\frac{P(A \mid B)P(B)}{P(A)} \tag{5-1}$$

推导过程如下：

$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}$$
$$P(B \mid A)=\frac{P(A \cap B)}{P(A)} \tag{5-2}$$

$$P(A \cap B)=P(A \mid B)P(B)=P(B \mid A)P(A) \tag{5-3}$$

其中，$P(A|B)/P(A)$称为可能性(likelihood)函数，这是一个调整因子。那么贝叶斯公式可以理解为

后验概率＝先验概率×可能性函数

首先，通过经验预估一个先验概率。然后通过实验数据进行不断的调整。如果$P(A|B)/P(A)>1$，那么意味着先验概率被增强，事件 B 发生的概率变大；如果 $P(A|B)/P(A)=1$，那么说明事件 A 和事件 B 是相互独立的，彼此没有影响；如果 $P(A|B)/P(A)<1$，那么意味着先验概率被削弱，事件 B 发生的概率变小。

如果 A、B 为两个独立的随机事件，那么

$$P(A \mid B)=P(A),P(B \mid A)=P(B),P(AB)=P(A)P(B)$$

如果事件 B 是由 n 个互不相容的事件 $B_1,B_2,\cdots,B_n$ 个事件组成($B_1,B_2,\cdots,B_n$ 是 B 的一个完备事件组)，那么 $P(B_i)>0,i=1,2,\cdots,n$，且 $B=\bigcup_{i=1}^{n}B_i$。而事件 $A\subset B$，那么 $P(A)=P(A\cap B_1)+P(A\cap B_2)+\cdots+P(A\cap B_n)$，其中，$P(A\cap B_i)=P(A|B_i)P(B_i)$。那么事件 A 的发生概率为

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A \mid B_i) \tag{5-4}$$

这个公式称为全概率公式。全概率公式的意义在于将事件 A 划分为比较简单的事件 AB_i(即 $A\cap B_i$)，然后通过加法公式和乘法公式计算出 A 的概率，事件 A 的发生有各种各样的原因，即 B_i，如果 A 是由某个原因引起的，那么全部原因 $B_i(i=1,2,\cdots,n)$引起 A 发生的概率的总和就是全概率公式。全概率公式可以看作从原因推出结果的公式，即每个原因都对结果有一定的作用。结果的发生与各种原因都有关系。

若 $P(B_i)>0,i=1,2,\cdots,n,P(A)>0$，那么在事件 A 发生的条件下，事件 B 发生的概率为

$$P(B_i \mid A)=\frac{P(B_i \cap A)}{P(A)}=\frac{P(B_i \mid A)P(A)}{P(A)}=\frac{P(B_i)P(A \mid B_i)}{P(A)}$$
$$=\frac{P(B_i)P(A \mid B_i)}{\sum_{i=1}^{n}P(B_i)P(A \mid B_i)} \tag{5-5}$$

这个公式就是著名的贝叶斯公式，该公式揭示了在事件 A 已经发生的条件下，可以用它来寻找导致 A 发生的各种原因的 B_i 的概率。

4) 贝叶斯判定准则

假设 $\Omega=\{C_1,C_2,\cdots,C_k\}$是 k 个不同类别的集合。现有样本集 $S=\{S_1,S_2,\cdots,S_m\}$，每个样本都是一个 n 维的特征向量 $\boldsymbol{X}=[X_1,X_2,\cdots,X_n]$，$P(\boldsymbol{X}|C_i)$是特征向量 $\boldsymbol{X}$ 在类别 C_i 状态下的条件概率，$P(C_i)$是类别 C_i 的先验概率。根据贝叶斯公式，后验概率

$P(C_i|\boldsymbol{X})$的计算公式为

$$P(C_i \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid C_i)P(C_i)}{P(\boldsymbol{X})} \tag{5-6}$$

其中，$P(\boldsymbol{X}) = \sum_{i=1}^{k} P(C_i)P(\boldsymbol{X} \mid C_i)$。

基于贝叶斯公式，可以计算出后验概率最大的 $P(C_i|\boldsymbol{X})$，作为样本 S_j（$1 \leqslant j \leqslant n$）的分类，这就是贝叶斯判定准则，它的形式化描述如下：

如果对于任意的 $i \neq j$，都有 $P(C_i|\boldsymbol{X}) > P(C_i|\boldsymbol{X})$成立，那么特征向量 $\boldsymbol{X}$ 对应的样本 S_j 被判断为类别 C_i。

5）极大后验假设

根据贝叶斯公式，可以得到一种计算后验概率的方法：在一定假设的条件下，根据先验概率和统计样本数据得到后验概率。

令 $P(c)$是假设 c 的先验概率，它表示 c 是正确假设的概率，$P(\boldsymbol{X})$表示的是训练样本 $\boldsymbol{X}$ 的先验概率，$P(\boldsymbol{X}|c)$表示在假设 c 正确的条件下，样本 $\boldsymbol{X}$ 发生或出现的概率，根据贝叶斯公式可得，后验概率的计算公式为

$$P(c \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid c)P(c)}{P(\boldsymbol{X})} \tag{5-7}$$

设 C 为类别集合，对于给定的未知样本 $\boldsymbol{X}$，通过计算找到可能性最大的假设 $c \in C$，具有最大可能性的假设或类别被称为极大后验假设（maximum a posteriori），记作 c_{map}：

$$c_{\text{map}} = \arg\max P(c \mid \boldsymbol{X}) = \arg\max \frac{P(\boldsymbol{X} \mid c)P(c)}{P(\boldsymbol{X})}, \quad c \in C \tag{5-8}$$

由于 $P(\boldsymbol{X})$与假设 c 无关，上式可以变为

$$c_{\text{map}} = \arg\max P(\boldsymbol{X} \mid c)P(c), \quad c \in C \tag{5-9}$$

6）朴素贝叶斯分类模型

朴素贝叶斯分类（naive Bayesian classification）是一种十分简单的分类算法。其本质思想是：对于给出的待分类项，求解在此项出现的条件下将其归属于每个类别的概率，概率最大的值对应的类别就是待分类项属于的类别。举个例子，正在使用的计算机突然出现了蓝屏，导致蓝屏的原因可能是计算机感染病毒，也可能是操作系统出现故障，但通常第一反应是计算机感染病毒了，因为计算机中毒导致蓝屏的现象比较常见。在没有其他可用信息的情况下，人们往往会选择条件概率最大的类别，这就是朴素贝叶斯分类的思想基础。

朴素贝叶斯分类是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入输出的联合概率分布；然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。朴素贝叶斯算法用于学习具有属于特定类的某些特征的对象的概率。简言之，它是一个概率分类器，之所以称它为“朴素”是因为它假设某个特征的出现与其他特征的出现是独立的。朴素贝叶斯分类器的结构如图 5-1 所示。

假设样本空间有 m 个类别$\{C_1, C_2, \cdots, C_m\}$，数据集有 n 个属性$A_1, A_2, \cdots, A_n$，给定