工业时间序列异常类型及案例分析

异常具有的一些特性

高纬度：数据集纬度高，数据间相互交织，人工检测不太可能。

最低假阳性：作为异常检测问题，我们不希望有过多的假阳性报警来干扰监控人员。

周期性：每小时/每天/每周/每月这样的周期性数据如果不妥善处理，某些周期性的行为可能误报为异常。实际数据中，每天固定时段的峰值数据相对于大部分采样点都可能被判定为异常，但实际为周期性正常现象。

趋势性：数据并不是均匀分布的，算法需要足够健壮来处理非均匀分布的数据集(增长性数据是一个普遍现象，如长期来看的股市指数等)。

数据集中少数的异常点会形成少量的聚类。

异常分类

按照表现形式：点异常（outlier），模式异常（outlier pattern）。

- 对于点异常（outlier，离群点），又有几种分类方式：加性异常点 (AO : addtionaloutlier)、革新异常点 (IO:innovativeoutlier)、水平漂移 (LS : levelshift) 和暂时变化(TC:temporary change)。
    - 加性异常点：孤立的异常点，并不波及到后面的观测值。
    - 革新异常点：通常涉及到时间序列的内在相关结构，往往成片出现。
    - 水平漂移异常点：加性异常点的一种特例，这种异常点变化较持久。
    - 暂时变化异常点：加性异常点与水平异常点的推广，其影响是指数衰减，这种影响会在以后慢慢消失。

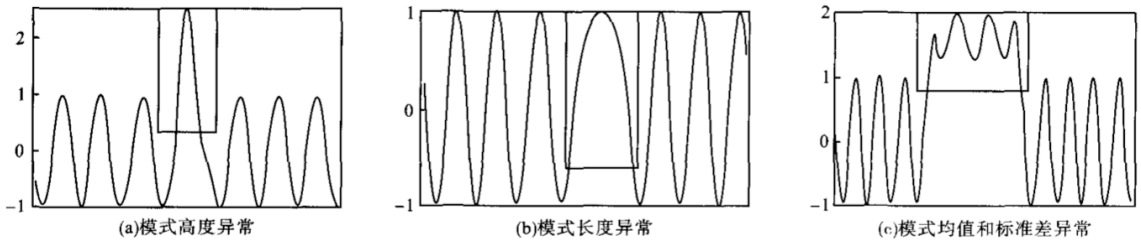


图 2 时间序列异常模式的三种表现形式

- 对于模式异常（outlier pattern）

> 时间序列的模式指时间序列的某种变化特征，它可以是时间序列离散化之后的符号，或者是时间序列的傅里叶变换系数等。
>
> **通过提取时间序列的模式，将时间序列变换到模式空间，就得到了时间序列的模式表示。**

    - 模式高度异常
    - 模式宽度异常
    - 模式均值和标准差异常

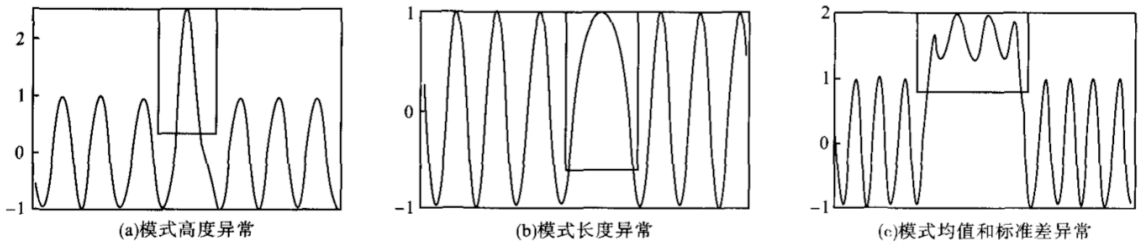


图 2 时间序列异常模式的三种表现形式

一、输电设备状态异常

输变电设备全景状态信息呈现来源多、信息异构、数量庞大、属性繁多等特点,其数据往往是不完整的、有噪声的和不一致的。状态量原始的数据质量往往不能满足后续状态评价模型的要求，因此，在状态评估或诊断分析之前进行数据清洗是必不可少的。数据清洗通过填充缺失值、平滑噪声数据和识别离群点来提高数据质量，有助于提高数据挖掘过程的准确率和效率。

异常表现形式

根据输变电设备的运行特点，状态数据中的异常通常表现为两种形式:

1. 可用于数据清洗的异常，即噪声点和缺失值；
    - 噪声点是指由于仪器异常或设备系统的扰动引起的严重偏离期望值的数据，这些数据不仅会影响模型拟合的精度，而且会导致后续状态评估出现偏差，引起误诊。
    - 缺失值是指由于传感器的短时失效、通信端口异常、记录失误等因素引起的数据中断，状态数据中存在的缺失值破坏了系统运行的连续性，不利于后续的状态评估和趋势检验。
2. 设备运行状态受到干扰而导致的数据异常。
3.
    - 设备在运行过程中会产生突发性故障、绝缘劣化等，这些常常会引起数据的水平迁移异常和趋势改变性异常，此类异常数据反映了设备运行工况的异常，不属于清洗范畴。

设备状态数据的时间序列中往往含有多个异常数据，修复所有的噪声点和缺失值是设备状态数据清洗的目标，同时也要实现突发性故障信息的有效获取，而不是作为异常数据剔除。

异常实例

文章中举出了一个变电站油中气体 CH_4 的在线监测数据，通过清洗之后发现 $t=50$ 处变压器出现了异常运行状态，需要运维人员密切关注。经过调试后发现实际情况是变压器低压侧上夹件内衬铁斜边与 C 相端处相碰形成了故障接地点，从而与原接地点形成了环流，使得变压器过热。

> 论文参阅：输电设备.pdf

二、设备工况

简述

随着全球经济快速发展，产品设计制造商和产品用户更加关注产品的维修保养服务，如何减少大型复杂设备的维修费用、缩短服务响应时间和提高设计生产质量是产品生命周期状态监测与运维服务关注的重点和难点。

异常分析

本文所指设备工况数据为设备监测系统实时监测设备状态并回传的数值型数据，是监测数据的一种，它描述了该设备在一定条件下的工作状态。复杂设备的工况数据一般有如下几个特点：

1. 维度高：不同工况数据为分析设备状态提供了不同的视角，即维度。因此维度高指一台设备的工况种类繁多，例如一台五十铃三桥46m泵车有270多个工况采集点，采集到的工况数据共同反映了该泵车的整体运行状态。
2. 数据量大：由于设备种类和数量多，每种设备的工况种类繁多，且监测系统按周期回传工况数据，使得一定时间内累计的设备工况数据量大。例如某工程机械制造企业生产的设备在线8万台，每日存储工况数据可达2亿条记录。
3. 数据类型复杂：不同工况数据的表现形式不同，包含布尔量，模拟量等各种类型，如紧急停止，高压启动等为布尔量；发动机转速，分动箱转速等为模拟量。
4. 采集周期不一致：采集周期为设备监测系统，回传工况数据的周期。由于工况数据特点和用户关注度的不同，不同工况数据设置的采集周期也不同。例如紧急停止工况数据采集周期不固定，当泵车紧急停止按钮按下时开始采集回传，而分动箱转速工况数据的采集周期则设置为固定的1h。此外，由于一些客观原因导致某些工况数据采集稀疏，数据质量较差，例如发动机转速工况数据每时每刻都会变化，但其采集周期为每小时一次，采集时间间隔久，丢失了很多时间序列信息。

异常数量较多的工况说明与其关联的一个或多个构件易磨损，是产生潜在故障的重大隐患工况。而异常数量较多的设备可分为两种情况：①设备某工况频繁出现异常，它可能反映了因磨损、疲劳等带来的潜在故障或用户频繁的不正当操作行为；②设备多种工况出现异常，它可能反映了由多工况共同影响的潜在故障。以上分析对于潜在故障的挖掘十分重要，因此异常检测的最后一步为将检测和合并后的异常数据按设备、工况和异常类型进行统计排序和分类展示，将出现异常数量较多的工况和设备异常数据提供给相关领域专家进行人工验证。

异常实例分析

本实验数据来自于某工程机械制造企业279台泵车6个月的200多种工况数据，共5283万个设备工况数据。每个工况数据都记录了设备工况单时刻的取值。

此文章定义了七种异常的基本类型，在系统界面图中可以看出监测结果。

表 1　七种特征提取方法

名称	介绍	适用数据类型	检测意义
最大值 Max	数据中最大可达值	模拟量	检测数值过高异常。如检测发动机转速过高情况等
最小值 Min	数据中最小可达值	模拟量	检测数值过低异常。如检测电流过小情况等
平均值 Avg	所有数据的平均值	布尔量和模拟量	检测数据集中程度异常。如检测日平均耗油量等
方差 Stdev	所有数据的方差值	布尔量和模拟量	检测数据变化程度异常。如检测倾角变化幅度异常等
下降率 Desc	后一个数据小于前一个数据的统计概率	模拟量	检测数据跳变程度和单调性异常

续表 1

升高率 Asc	后一个数据大于前一个数据的统计概率	模拟量	检测数据跳变程度和单调性异常。如检测工作时长等
采集频率 Freq	每秒采样个数	所有类型	检测采集频率和其潜在异常现象，如检测频繁报警情况、传感器失常等

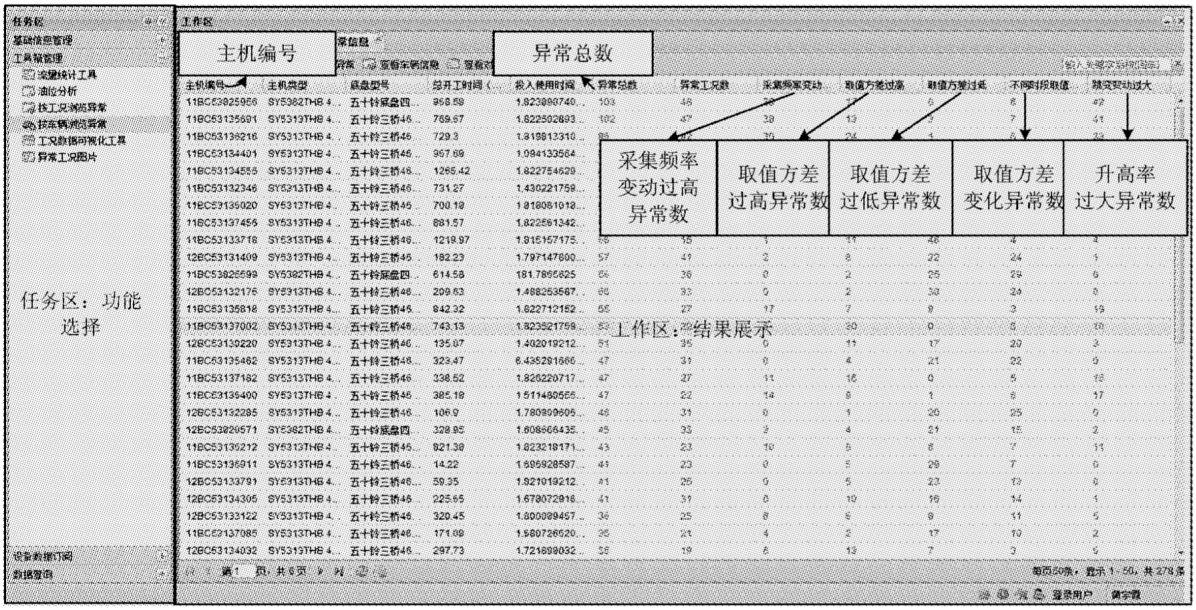
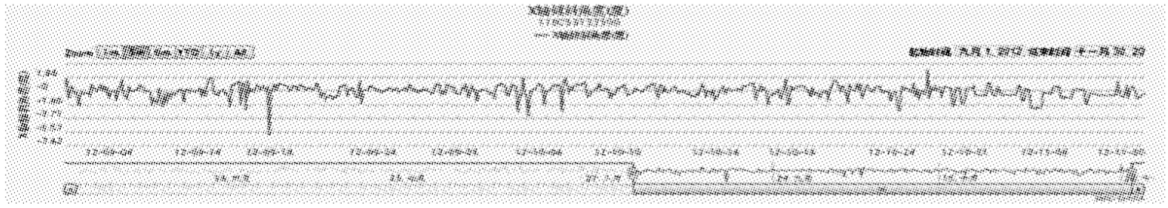


图6　系统界面图

a x轴倾斜角度工况异常数据



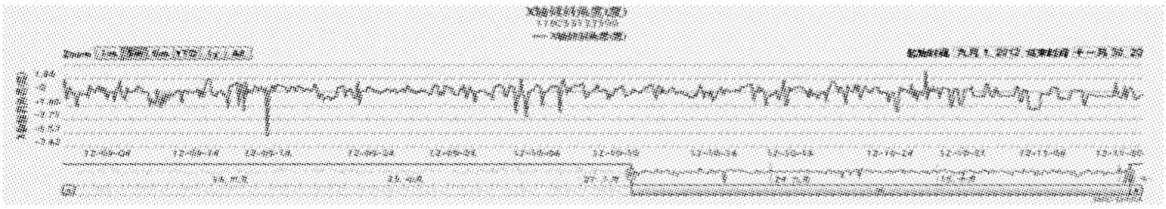
b x轴倾斜角度工况正常数据

图7　x轴倾斜角度无变化异常图

经领域专家的分析和实验检测出的异常数据，对识别用户异常行为、发现安全隐患和促进质量改进等方面十分有帮助，从而证明了所检测的异常数据的价值和异常检测方法的有效性。

例如通过取值方差过小异常检测，本文发现 x 轴倾斜角度工况出现异常情况。经异常检测方法所得的该工况正常取值方差特征空间范围应在0.3~0.5之间，因此，如图7b中的正常工况数据所示：编号为11BC531133590泵车的x轴倾斜角度值会在 0 值范围内小幅度波动。但图7a中编号为11BC531132319泵车的x轴倾斜角度一直为0，是
异常数据。

a x轴倾斜角度工况异常数据



b x轴倾斜角度工况正常数据

图7　x轴倾斜角度无变化异常图

检测出的异常数据经由领域专家分析验证表明，该工况数据为反映泵车离地倾斜角度而设计，其正常的表现形式确实与分析所得结果一致，其值不断波动的原因是当泵车倾斜角度超过一定值时，泵车会自动进行调整，保持平衡。

经调查发现，异常现象是由于客户希望泵车长期运行，自行拔出倾角传感器造成的，属于不正当异常操作行为。这种行为会使得泵车防倾覆功能失灵，产生巨大的安全隐患，若没有及时发现，则后果不堪设想。

在促进质量改进方面，通过升高率过小异常检测，发现主机工作时间异常情况。经人工观察验证，检测出的异常数据有很多跳变，确实与主机工作时间正常工况数据的单调递增特征相悖。

经了解，确定其原因为泵车错误回传压路机信源，目前公司正在解决该问题。以上示例充分表明了异常检测的重要价值和本文所提出的异常检测方法的可行性和有效性。

> 工况设备.pdf

三、股票市场

简述

对于股票数据，面对大量的数据。如何寻找对投资者有利的信息，为了有良好的预期，本文尝试了一些新的挖掘方法。统计学家对于数据挖掘研究的重视和投入近年来显得根突出。同时很多国内外的学者都把重点放到了时间序列异常点的挖掘上，对于证券研究．一般是通过找出股票日线序列中的异常点来构筑投资策略以提高投资的成功率，避免盲目性，**同时它也可以用于改进模型，提高预测精度。**

实验数据

实验中采用上证指数(INDEXSH)在1998年1月1日至2001年12月31日期间内日收益率作为实验数据，实验数据来自于雅虎。

异常原因分析

在本论文中，利用主曲线算法从数据中分别得到异常点集合后，然后再进行相应的处理，并结合当时的宏观政策与事件进行比较，通过比对可以发现造成异常收益的原因。

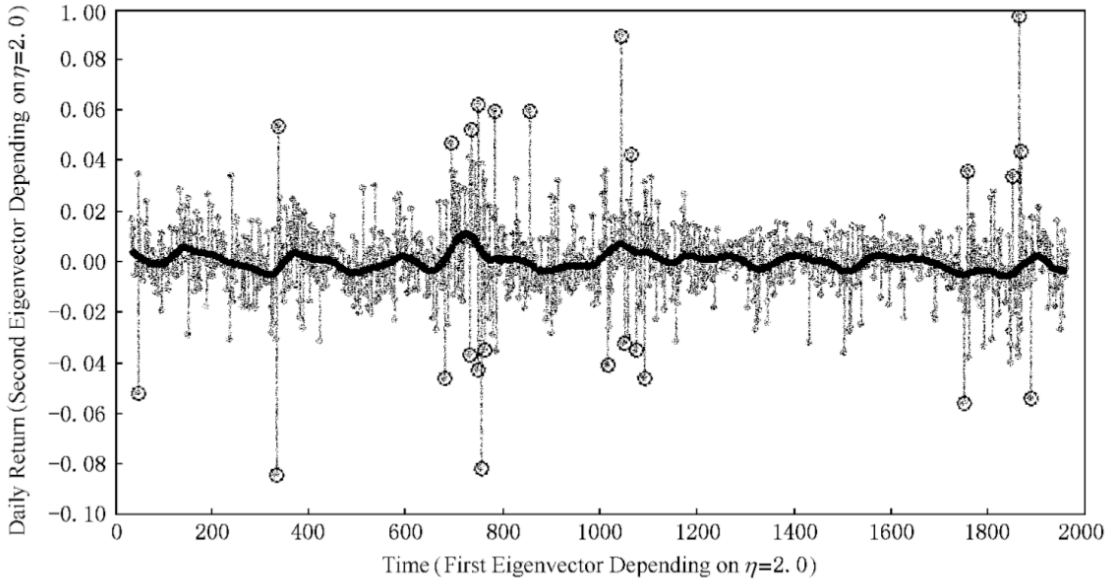


Fig. 3 The outliers detection result by the principal curve to R_η.

图 3 主曲线对 R_η 的异常检测结果

从图表中可知，在11次异常收益中，只有1次是有政治事件引起的，3次于经济事件相关，而7次则是由股票市场的宏观调控政策引起的，这说明，虽然宏观政治因素和经济因素也能导致股票的异常收益，但政府对股市的宏观调控政策是导致上证指数大盘发生异常的主要原因。

Table 1 The External Events of INDEXSH

表 1 上海大盘的异常事件检测结果

No.	Time	Duration				Events
		1	5	20	60	
1	19980113	Y				The Currency of Southeast Asian countries was impacted.
2	19980612-19980915	Y	Y	Y	Y	① The impact of Asia Financial Crisis on Southeast Asia; ② An unprecedented big flood occurred in China on Aug.
3	19990510	Y				US-led NATO attacked Chinese Embassy in Yugoslavia On May 9.
4	19990511-19990629	Y	Y	Y	Y	The six policies of" invigorating the market" caused the 5.19 market.
5	19990701-19990810	Y	Y	Y	Y	" Securities Act" formally took effect from 7.1.
6	20000104-20000111	Y	Y			General Office of the State Council announced the notice about several suggestions of setting up the risk investment mechanism.
7	20000125-20000316	Y	Y			The secondary market rationed new shares and stocks were pledged.
8	20001124	Y				China Securities Regulatory Commission issued notice on strengthening market supervision and beating market manipulation behavior; meanwhile nine inspection bureaus were to be established and 236 illegal cases in violation of rules and regulations were to be investigated, etc.
9	20010724-20010801	Y	Y		Y	The People's Bank of China investigated the funds in violation of rules and regulations.
10	20010918-20011022	Y	Y	Y	Y	① The reduction of state-owned stocks; ② Yinchuan Guangxia fake incident; ③ The 911 incident occurred in U.S.A.
11	20011023-20011107	Y	Y			The reduction of state-owned stocks was suspended.

Note:" Y" means there is an abnormal return in one of the daily, 5, 20 and 60 aggregate days' return sets. The announcements come from http://www.csrc.gov.cn/, the website of Chinese Securities Regulatory Commission.

> 论文参阅：金融数据.pdf

四、海洋数据

简述

本文的作者提供了一种基于时间序列异常检测技术的海洋信息管理系统，该系统主要包括数据采集，数据预处理，异常检测，数据存储，数据管理和展示等模块。海洋数据质量直接影响海洋信息管理系统的科学性，由于海洋信息数据库数据量太大，往往易受噪声，丢失数据和不一致数据的侵扰，因此该方法提供了一套基于累计变化量的时间序列异常检测方案，能够有效的检测出海洋数据中的异常点，再选择适当的修正方法对海洋数据中的异常点进行修正，从而建立一套完善的海洋信息管理系统，有效的对海洋数据进行管理，为我国的数字海洋建设提供有力支持。

异常原因分析

在这个系统中，数据采集模块包括气象传感器，水文传感器和生物传感器，气象传感器采集气象类数据，包括风速风向，气温，降雨量和雾量等数据；水文传感器采集数据包括水温，盐度，海流，波浪，潮位，含沙量和悬沙等；生物传感器采集浮游动物，浮游植物和底栖生物等数据。

通过分析，发现海洋信息管理系统中的元数据存在以下问题。

1. 数据库在某些字段上存在着空值，所以需要对这些数据进行一些转换和集成工作，对空值字段需进行数据的智能填充。
2. 各个站点关于台站信息的数据在结构上基本相同，但在数据的完整性和一致性上很差。
3. 来自不同数据表的同类数据，使用不同的数据类型表示，有的是字符型，有的是日期型。
4. 各台站的海洋数据中或多或少的含有噪声数据，在装入数据库前必须进行清洗。

综上所述，海洋系统的原始数据存在着数据不一致性，数据残缺，数据冗余等情况。

异常数据的产生原因很多，主要有以下分类：

1. 传感器暂时故障导致值缺失，之后时间的数据覆盖掉缺失段的数据，导致周期趋势不存在，被检测出异常。
2. 传感器故障导致数据错误，这种错误通常反应都是数据长时间出现极值。
3. 人工因素导致传感器摆放位置错误，记录的数据与理想状态不符合。

> 论文参阅：海洋数据.pdf

五、NDVI时序数据

NDVI 是监测地表植被活动简单、有效的度量指标。目前基于 NOAA/AVHRR、MODIS、SPOT/VEGETATION 等卫星遥感数据得到的 NDVI 时序数据已经在植被动态变化监测、宏观植被覆盖分类和植物生物物理参数反演方面得到了广泛应用。应用中低分辨率传感器对地面进行观测时，太阳光照角度、观测角度以及云的覆盖状况随时间变化很大，因此得到的观测值包含了大量的噪声。尽管观测数据经过严格的预处理，并采用最大值合成法(Maximum Value Composite, MVC)或者限制视角的最大值合成法 (Constrained-View Angle Maximum Value Composite, CVMVC)将多天的 NDVI 数据进行合成，但是得到的产品仍然受云、气溶胶及水汽等的影响。尤其是当合成期内一直有云存在的情况下，云成为对 NDVI 产品影响最大的噪声。这些噪声的存在使得合成后的NDVI随时间变化呈无规律状态，相邻值高低变化没有规律，难以进行应用。

实验数据

采用的实验数据为山东省一年内的MOD13A2产品。该产品共有23幅图像，缩放尺度为10000，有效值范围为 -2000 ~ 10000，填充值为 -3000。

山东省典型土地利用类型为农田和林地, 在选择典型地物的纯净像元时, 参考了1:25万土地覆盖遥感调查与监测数据库。

异常实例

- 在检测过程中发现在农作物序列的第8个合成期NDVI有一个小的下降，这不太符合健康作物正常的生长规律，有可能该时期作物生长受到了环境的胁迫, 也有可能是受薄云的影响。
- SPLINE 插值后的图像有亮斑出现，其像元值超出了 NDVI 的有效范围。经过对像元时间序列分析对比，发现这种情况一般出现在时间序列的两端(冬季)有连续 2 个以上的像元都受到云或冰雪影响的时候。

> 论文参阅：http://www.gtzyyg.com/article/2011/1001-070X-23-1-33.html

六、网络异常流量监测

DoS(denial of service, 拒绝服务) 攻击是对网络服务有效性的一种破坏, 使受害主机或网络不能及时接收并处理外界请求, 或无法及时回应外界请求, 从而不能提供给合法用户正常的服务, 形成拒绝服务。DDoS 攻击就是利用足够 数量的傀儡机产生数目巨大的攻击数据包对一个或多个目标实施 DoS攻击，耗尽受害端的资源，使受害主机丧失提供正常网络服务的能力。DDoS 攻击已经是当前网络安全最严重的威胁之一，是对网络可用性的挑战。反弹攻击和 IP 源地址伪造技术的使用使得攻击更加难以察觉。就目前的网络状况而言，世界的每一个角落都有可能受到DDoS 攻击，但是只要能够尽可能检测到这种攻击并且作出反应，损失就能够减到最小程度。因此，DDoS攻击检测方法的研究一直受到关注。

由于大部分用户浏览行为的统计特征(点击速度、浏览的内容或请求对象、浏览时间、浏览过程等)具有一定的相似性，因此，我们可以把这种统计特征看作是用户正常的、合法的行为。相对于正常的普通用户来说，攻击者发出的攻击流通常具有以下特点。

- DDoS攻击发生时具有特定的模式，比如网络中会出现流量突增并超过正常工作时极限流量的现象，如果当前网络流量超过了阈值则说明可能发生了 DDoS 攻击。
- DDoS 攻击发生具有一些特定的前置特征，例如每次攻击发生前夕，攻击者要解析受害者的主机名，因此网络中就会出现大量的地址解析请求，如果解析后发现同一个主机名称出现过多的话，则可能发生攻击。
- 某些特定的攻击会有明显的模式，例如 TFN2K 攻击时会发送一些数据段只包含文字和数字字符的数据包，其他一些攻击则发送数据段只包含二进制字符串和 high-bit 字符的数据包。

> 论文参阅：DDoS文件夹