该技术报告的修订版将出现在2009年9月的ACM计算调查中。

异常检测：一项调查

异常检测在不同的研究领域和应用领域的研究中都是一项很重要的问题。针对某些应用领域专门开发了许多异常检测技术，而其他的则更为通用。这项调查尝试提供一个结构和全面的观点在异常检测研究领域当中。我们根据每种技术采用的基本方法将现有技术分为不同的类别。对于每种类别我们定义了关键的假设，这些假设被这些技术用来区分正常和异常行为。当申请已有的技术在常见的领域当中，这些假设可以当做一个指导原则来评估该技术在这个领域内的有效性。（在将给定技术应用于特定领域时，这些假设可用作评估该技术在该领域中的有效性的指导原则）。对于每一个类别，我们提供基本的异常检测技术，然后说明该类别中不同的现有技术是基本技术的变体。该模板提供了对每种类别技术的更简单明了的理解。进一步，对于每个类型而言，我们定义了每一个技术在该类型中的优点以及缺点。我们还讨论了这些技术的计算复杂性，因为它是实际应用领域中的重要问题。我们希望这项调查可以更好地理解针对该主题进行研究的不同方向，以及如何将某一领域中开发的技术应用于并非其初衷的领域。

介绍

异常检测是发现有问题的数据模式，这些数据模式和预期的行为不相符合。通常这些不合格的模式通常被称为不同应用领域中的异常，离群值，不一致的观察结果，异常，出人意料，特殊性或污染物。其中，异常和离群值是异常检测中最常用的两个术语；有时可以互换。异常检测可被广泛应用在各种不同的应用中比如说信用卡诈骗、保险或是卫生保健、入侵检测以实现网络安全，安全关键系统中的故障检测，以及对敌方活动的军事监视。（异常检测可广泛用于各种应用程序中，例如信用卡欺诈检测，保险或医疗保健，网络安全入侵检测，安全关键系统中的故障检测以及对敌方活动的军事监视。）

异常检测重要性归因于以下事实：数据异常会转换为多种应用程序领域中的重要（且通常是关键的）可操作信息。例如，计算机网络中的异常交通模式通常意味着有骇客电脑正在发送敏感数据到一个未经授权的目的地。一个异常的RMI图片可能表明当下有一个恶性肿瘤。在信用卡交易数据的异常能够表明是否发生了信用卡或者身份盗窃或者来自航天传感器的异常读数可能表示太飞船某些组件故障。

早在19世纪，统计界就对检测数据中的异常值或异常进行了研究。随着时间的推移（over time），各种各样的异常检测技术被几个研究社区开发出来。许多异常检测技术被特别发展对于某些应用领域，而其他的一些更为通用一点。本次调查试图提供一个异常检测研究的结构化的和全局化的视角。我们希望这能更好地帮助理解在这个主题上已经进行了不同方向的研究，以及这项技术在一个领域如何发展以及能够被应用于不打算以其开头的领域中。（在一个领域中开发的技术如何应用于不打算以其开始的领域。）

* 1. 什么是异常

异常是数据的一种模式，这种模式是他不舍和一个已经定义好的正常行为的概念。图1展示的在一个简单的二维数据集中的异常。这些数据有两个正常的区域，N1和N2，因为大部分的观察点豆斋这两个区域里。充分远离这两个区域的点，比如点Q1，Q2，以及落在区域Q3中的点，就是异常点。

处于各种原因，异常也许会包括在数据当中，比如说恶意的运动，银行卡欺诈，网络入侵，恐怖活动或者是暂停一个系统，但是所有的这些原因都有一个共同的特征，那就是数据 分析人员对此很感兴趣。异常的趣味性或与现实生活的相关性是异常检测的关键特征。

异常检测与噪声消除和噪声适应有关，但与噪声消除和噪声适应不同，两者均处理数据中不需要的噪声。噪声可以被定义为数据中的一种现象，它对数据分析没有兴趣，但是在操作上会阻碍数据分析。去噪是因为我们需要移除不想要的目标，在对数据执行任何数据分析之前。噪声调节是指对统计模型估计值进行免疫以防止异常观察。

另一个异常检测相关话题是新奇检测（novelty detection），这个主要检测先前未观察到的（新兴的）在数据中的模式，比如说：一个新的小组中的新的讨论话题。新颖模式和异常之间的区别在于，新颖模式通常在被检测到之后被合并到正常模型中。

应该注意的是，针对上述相关问题的解决方案通常用于异常检测，反之亦然，因此在本综述中也将对此进行讨论。

1.2 挑战

在抽象水兵上将，异常被定义为一个不同于正常行为的模式。

13 总结和今后的工作

在本次调查中，我们讨论了文献中提出的异常检测问题的不同的方式，（并试图提供一个全局的）并试图提供有关各种技术的大量文献的概述。对于每一个领域的异常检测技术，我们定义了关于正常和异常数据概念的唯一假设。