通用可扩展的时间序列异常自动检测框架

摘要

该论文针对大规模时间序列数据提出了一种既通用又可以扩展的异常自动检测的框架。早期的异常检测在维护用户数据的连贯性以及保护企业免受恶意攻击方面扮演了很关键的角色。目前最先进的异常检测方法存在可扩展性、用例限制、使用困难和大量误报等问题。我们在雅虎的EGADS系统，使用了一组带有异常检测过滤层的异常检测和预测模型，以便在时间序列上进行准确和可扩展的异常检测。我们将我们的方法和其他具有不同时间序列特征的实时和合成数据异常检测系统进行了比较。我们发现我们的框架允许在精度和召回率方面为各种用例提高50-60%。数据和框架我们都已开源。特别地，开源的数据为这类异常检测工作提供了一个标准参考，这也是其开创的一个先河。

介绍

尽管计算机硬件和软件的快速发展带来了强大的应用程序，数以百计的软件bug和硬件错误在大规模的计算机集群中持续发生影响着用户的体验进而影响收入。不间断系统有严格的正常运行时间要求，对这些系统的持续监控至关重要。从数据分析的角度看，这意味着不间断地监控很多的时间序列数据从而发现隐藏的失败或异常。因为问题规模很大，人工监控这些数据从实际来讲不可行，这也导致我们使用机器学习和数据挖掘的方法进行自动异常检测。