**吕梦 19291128 M3阅读笔记**

**选文标题：NETRCA: AN EFFECTIVE NETWORK FAULT CAUSE LOCALIZATION ALGORITHM**

**作者：***Chaoli Zhang*1 *, Zhiqiang Zhou*1 *, Yingying Zhang*2 *, Linxiao Yang*1 *, Kai He*3 *, Qingsong Wen*3*, Liang Sun*3

**一、文章内容**

1.摘要：网络的运行和维护很重要的一部分是定位网络故障的根本原因。但因为复杂的网络架构和无线环境、有限的标记数据，大大增加了准确定位真正的根本原因的难度系数。为了解决这个问题，文章提出了一种新的算法：NetRCA。首先，考虑时间、方向、属性和交互特征，从原始数据中提取有效的衍生特征。然后，采用多元时间序列相似性和标签传播，从已标记和未标记的数据中生成新的训练数据，增加标记样本的数量。接着，设计出一个集成模型，这个模型将XGBoost、规则集学习、属性模型和图算法相结合，可以充分利用所有的数据信息，同时提高性能。最后，在ICASSP 2022 AIOps挑战赛的真实数据集上进行实验和分析，证明本文提出的方法的优点。

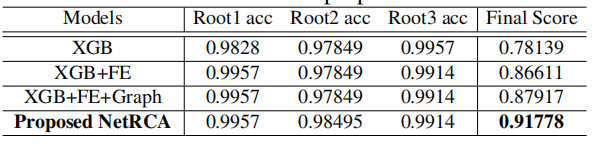
2.文章论证叙述清晰，由介绍、提出网络RCA框架、实验和讨论以及最后的总结组成。

通过介绍，我们可知论文的背景是随着网络规模和复杂性的增加，对自动智能的根本原因分析算法和工具的需求量大大增加。但因为三个关键障碍，现有的故障定位方案仍然难以解决复杂的5G网络。第一个障碍是网络深度的增加使得原因归因变得困难；第二个是缺乏足够的已知标签；第三个是难以提取节点关系。所以文章提出了一种有效的无线网络故障原因定位算法NetRCA来解决这些挑战。它由三部分组成：特征工程、数据增强和模型集成。同时由于标记数据在实践中有限、存在大量的未标记数据，文章提出了一种新的方法来进行数据增强来生成标记数据。最后，利用模型集成将根本原因定位作为一个分类问题，该模型集成不仅采用XGBoost 获得强基线，还利用规则集学习、归因模型和图算法，利用因果关系图进一步提高性能。

提出网络RCA框架通过框架概述先进行总结，然后依照特征工程、数据增强和模型集成这三部分分别进行介绍。



在实验和讨论中总结并讨论了NetRCA在ICASSP2022AIOps挑战数据集上的性能。在性能比较和消融术研究中，通过比较基线XGBoost模型、XGBoost模型特征(XGB+FE)生成特征工程、XGB+FE和图算法(XGB+FE+图)，和NetRCA算法，绘制出表1



从而从表中得出以下结论：（1）所有的模型，甚至是基本的XGB模型，都可以在训练数据中获得极好的准确性。文中提出的NetRCA算法可以防止过拟合，并给我们一个更棒的解决方案。（2）XGB+FE模型在训练和验证集上都显著优于基本的XGB。通过深入研究背景，从时间特征、方向相关特征、属性特征和交互特征中提取有效的信息，我们能够获得一个完整的视角，并发现一些潜在的规则。（3）虽然在训练集上没有明显的改善，但结合图模型 可以使最终的提交分数提高1%以上。原因可能在于，图模型可以帮助我们更好地捕捉这些特征之间的因果关系。（4）规则集学习和归因模型可以减少这些特征之间的相互影响，并进一步提高最终得分。

3.总结：在最后的总结中，我们知道本文提出了一种新的NetRCA算法来定位网络故障的根本原因。 除了精心设计的特征工程外，坐着的算法通过数据增强来生成新的训练数据，从而解决标记样本的缺乏。另外，作者还设计了一种集成方法，有效地结合了不同的模型，对网络故障进行准确可靠的因果推理。

**二、个人理解：**

通过结合其他相关资料的学习和拓展，我了解到：

在实际中，网络设备经常发生电力、链接等故障，故障时链路上的设备不断上报大量的告警数据，这些告警让运维人员应接不暇。跨域的告警分析、根本原因定位难度大，对运维人员技术要求高，另外属于同一故障的多个告警，让运维效率进一步下降。使用根本原因事件识别，可基于实时的告警流、拓扑数据，聚合故障相关事件、快速地识别故障并准确定位故障根因。

现有的一些根因检测算法大致包括时序算法、无监督方法、贝叶斯网络和可解释方法。

根因分类则是，已知根因可能的若干类型，利用模型判断最可能的一个或多个根因类型，通常是有监督训练场景下的。也是现在绝大多数已经发布的所谓“AI根因定位”产品可以应对、智能运维赛事中寻求解决的根因定位问题。

但是无论是使用XGBoost这样的分类模型，或是基于相似度的聚类模型，最核心的目标都是通过合理并且全面的特征工程，最终提升最终根因类型判别的准确度。但这类问题也存在非常明细的局限性，比如：

（1）难以处理历史上从未出现过的全新根因类型；

（2）标注样本量少，需要介入数据增强手段比如主动学习、GAN样本生成等；

（3）异常场景复杂导致分类类目多，效果差等。

那么，这些算法要么依赖大量充分的标签数据，要么缺乏对特征之间关联的建模，要么受限于计算效率难以大规模应用，可解释性的方法也是浅尝即止，不能很好地应对上面提到的诸多挑战。所以ICASSP这篇论文里面提到的新的算法框架能够很好的解决这一系列相关问题。

论文中提出的NetRCA框架主要包括了特征工程模块、数据增强模块和集成模型的方式，这些技术都有单独在实际应用中使用，各自发挥着重要的作用。但作者创新性的将它们结合起来后，新的算法框架产生了奇妙的功能，解决了复杂棘手的问题。

创新后的框架将时序特征、天线方向特征、归因特征和交叉特征等几大类特征结合，通过多时间序列相似性相关的技术，先计算了已打标记数据与未打标记数据间的相似程度，然后再根据相似性，给未打标记数据加标签，从而达到扩充训练所用的已打标记数据集的目的。最终再拿到XGBoost + 规则设定模型 + 归因模型 + 因果图的集成模型中进行训练，并将上述模型所产出的综合结论作为最终的推断结果。

我觉得，这个思想方法很值得我们学习，不一定必须绞尽脑汁设计出从未出现过的解决方法，而是善于运用已有的知识和方法，创造性的将其结合起来，各自发挥出重要作用。在这个过程中，对技术精准充足的理解和创新性的思维，以及科学的实验验证，才能产生实用的这个算法框架。