**论文阅读笔记**

19211112马巍孺

阅读论文的题目：NETRCA: AN EFFECTIVE NETWORK FAULT CAUSE LOCALIZATION ALGORITHM

目录如下：

目录

[**一、论文背景** 1](#_Toc104321266)

[**二、论文总体介绍** 2](#_Toc104321267)

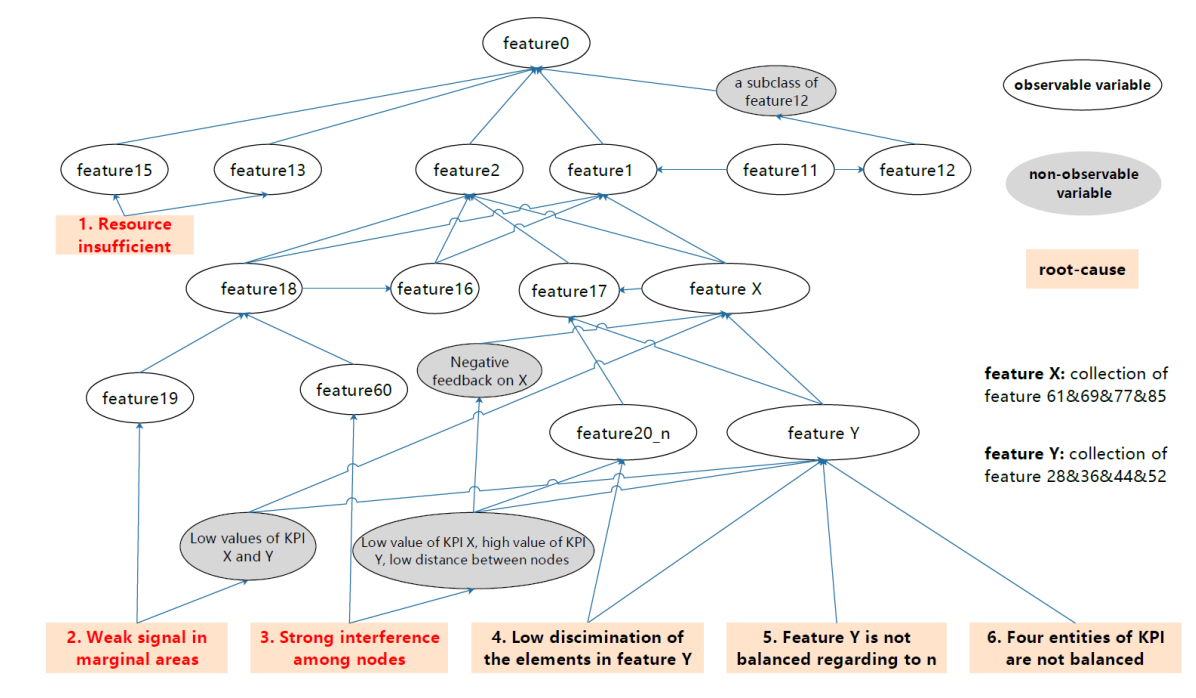
[**三、论文细节研读** 2](#_Toc104321268)

[**四、总结与体会** 8](#_Toc104321269)

**一、论文背景**

本篇论文是针对“ICASSP 2022 5G 网络故障根因定位挑战赛”而写的，这个挑战赛的背景是：5G无线网络故障根因定位是网络运维中的一个重要环节，通过快速且准确地判断网络故障的根因，技术人员可以及时采取措施对网络进行修复。然而，现在经常受困于复杂的无线通信环境和网络部署结构，且存在网络故障样本数少、不同的场景下故障表征差异性大等问题。如何充分利用领域知识和一小部分标定数据，使用统计学习和因果推断技术，快速准确地定位故障的根因，是网络运维面临的巨大挑战。

在电信网络中，低于预期的特征0的性能是客户关注的问题。本次大赛主要集中在 feature0 性能不佳的根本原因分析上。当其值小于200时，我们需要分析影响因素并找到原因。因果关系如下图所示。



每个功能都可以解释为电信网络中的关键绩效指标（KPI）。这些 KPI 的值随时间而变化，并相互影响。根据下图中提供的关系，需要推断 feature0 值较低的根本原因。

**二、论文总体介绍**

这篇文章的题目是：NETRCA: AN EFFECTIVE NETWORK FAULT CAUSE LOCALIZATION ALGORITHM（NETRCA：一种有效的网络故障原因定位算法），作者为：Chaoli Zhang， Zhiqiang Zhou，Yingying Zhang，Linxiao Yang，Kai He，Qingsong Wen，Liang Sun。

文章包括以下个部分：分别是摘要、 INTRODUCTION（简介），PROPOSED NETWORK RCA FRAMEWORK（该部分分模块介绍了算法的框架和数学原理），EXPERIMENTS AND DISCUSSIONS（该部分进行实验测试并得到了一些结论）， CONCLUSIONS和REFERENCES。

**三、论文细节研读**

**摘要部分：**

首先给出了问题的背景：定位网络故障的根本原因对于网络运行和维护至关重要。然而，由于复杂的网络架构和无线环境，以及有限的标记数据，准确定位真正的根本原因具有挑战性。

然后开始介绍本文的算法：本文提出了一种新的算法NetRCA来解决这个问题。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。其次，论文采用多元时间序列相似性和标签传播，从标记和未标记的数据中生成新的训练数据，以克服标记样本的不足。第三，论文设计了一个集成模型，该模型将XGBoost、规则集学习、属性模型和图算法相结合，以充分利用所有数据信息并提高性能。

最后，在ICASSP 2022 AIOps挑战赛的真实数据集上进行了实验和分析，以证明论文的方法的优越性和有效性。

**1.INTRODUCTION：**

在该部分中，首先介绍了一些背景的内容。网络规模和复杂性的增加需要自动和智能的根本原因分析算法和工具。由于网络类型的多样性和多变量时间序列数据的复杂关键性能指标（又叫KPI）模式，开发鲁棒可靠的故障定位解决方案具有挑战性，并受到了大量研究关注。

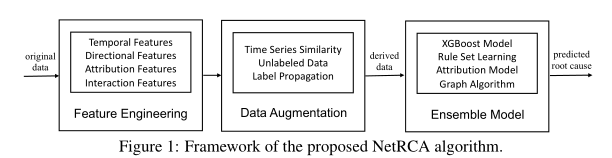
然后列出了一些在参考文献中提出的解决方法。比如基于无监督自组织映射的自动诊断系统，自适应根本原因分析的自动故障检测和诊断解决方案（该解决方案使用测量值和其他网络数据以及贝叶斯网络理论来执行自动循证RCA），自动诊断算法（该算法用于分析多个度量的时间演化，并在存在故障的情况下搜索潜在的相互依赖性），通过稳健时间序列分析和层次贝叶斯网络设计的根本原因分析系统，应用多种全局和局部可解释性方法，主要目的是通过识别有助于决策的重要特征，在网络中进行根本原因分析等。

接下来论文分析了现存的障碍。这些障碍导致现有的故障定位方案对于复杂的5G网络仍然很困难。第一个挑战是，网络深度的增加可能会通过从源节点到根节点的因果路径传播错误，从而使准确的根本原因归因变得困难。第二个挑战是缺乏足够的已知标签。在缩小可能的根本原因候选范围之前，通常必须深入研究KPI、服务日志和通信细节。最后，与每个网络节点关联的时间序列数据是多变量的。它们通常是复杂的模式子，具有相互依赖性和噪声，导致难以提取节点关系。

最后，介绍了本文的主要贡献。本文提出了一种有效的无线网络故障原因定位算法，称为NetRCA。NetRCA由三个主要组成部分，包括特征工程、数据扩充和模型集成。在特征工程中，为时间序列数据和无线方向相关特征设计了特征。由于在实际应用中，标记数据往往是有限的，而存在大量的未标记数据，因此论文提出了新的方法来执行数据扩充以生成标记数据。最后，论文使用模型集成将根本原因定位视为一个分类问题，该模型集成不仅采用XGBoost来获得基线，还利用规则集学习、属性模型和图算法，利用因果关系图进一步提高性能。除了具有较高的预测精度外，由于采用了基于规则的模型，NetRCA模型能够输出可解释的结果，这有助于了解根本原因如何影响系统。

**2. PROPOSED NETWORK RCA FRAMEWORK：**

**2.1. Framework Overview（框架概述）**



图一为算法的框架：算法的三个部分将在以下三小结中依次介绍。

**2.2. Feature Engineering（特征工程）**

由于每个样本中的时间戳数量不同，使用所有时间戳直接训练模型可能会导致偏差，即模型可能倾向于使用更多的时间戳索引来聚焦样本。因此，根据从每个样本中提取的特征来训练模型。生成的特征可以大致分为四类：时间特征、方向相关特征、属性特征和交互特征。

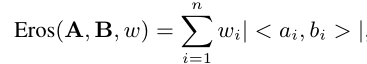
模型中使用的一些时间特征是基于数据统计的，其中每个时间戳中的数据都假设是独立的。5G网络中采用了多天线和波束成形来提高性能（多天线和波束赋形都是无线通信中的一些技术）。波束成形的方向和每个节点之间的距离在网络性能中起着至关重要的作用，这对于根源定位也很重要。我们认为特征20是检测根本原因2和3的模型中的一个重要特征。由于特征20给出了映射到4×8位置矩阵的每个节点从0到31的索引，因此首先将每个节点的索引转换为二维坐标，然后通过欧氏距离测量每对节点之间的距离。然后推导了用于捕捉特征X和Y之间相互关系的特征，分别包括特征61/69/77/85和特征28/36/44/52，以便进一步改进。最后，从每个时间片样本的距离分布中总结统计特征（如均值、方差、分位数等），用于模型训练。

根据因果图推导出除特征0之外的所有节点的属性特征。正如问题描述中所述，这些根本原因最终导致功能0的值较低。真正的根本原因及其后代将比其他因素对功能0的当前的值发挥更重要的作用。因此，生成了一个新的特征作为预测特征0上每个特征重要性得分的估计值。生成了X和Y的二阶交互特征。当特征X等于某些未知因素的特征Y比率时（X与Y成比例），生成特征X基于Y来衡量这些未知因素的影响。同时，首先根据问题描述将特征分组为X和Y对。对于每一对，计算X相对Y的比率。最后，像对时间特征所做的那样计算这些比率的统计信息。

**2.3. Data Augmentation（数据扩充）**

多元时间序列的相似性：

论文运用Eros算法（该算法引自其他论文）去计算多元时间序列的相似性，并扩展了Frobenius范数（范数为泛函分析领域的知识点，范数本身就是一种结构，拥有了这种结构，集合就具有特殊之处了）。具体需要用到奇异值分解等内容，最终得到的相似性为（w为权重向量）：



数据/标签扩充：

数据扩充对于从时间序列中学习很重要，因为标记的数据通常是有限的。会发现一半以上的训练数据是未标记的，仅仅删除这些数据就会丢失很多有价值的信息。使用Eros能够测量任意两个训练样本之间的相似性，这两个样本都是多变量时间序列。这样可以通过从那些与标记数据具有高度相似性的标记数据中选择样本来丰富训练集，并根据与这些样本相似的训练样本的真正根本原因来标记它们。针对每种类型的根本原因分别执行此过程，以提高计算效率。

另一个重要的增强是传播共享相似时间戳的训练样本的根本原因标签。这改进了对测试数据集的多个根本原因的预测，此处论文中对竞赛提到的一些根本原因进行了分析，即假设根本原因1更有可能独立于其余根本原因。因此，根据时间戳对齐了所有训练样本，并将其真实标签扩充为所有根本原因标签的联合集。

**2.4. Ensemble Model（集合模型）**

NetRCA采用集成模型预测根本原因，该模型应用XGBoost获得初始结果，然后结合规则集学习、归因模型和图形算法对结果进行细化，以获得最终结果。细节如下：

通过XGBoost进行根本原因分类：

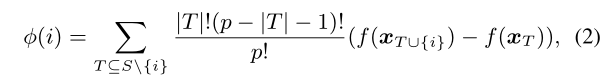
在解决方案中，论文将找到正确的根本原因视为一个分类问题。具体来说，采用XGBoost作为基础模型，因为它具有良好的性能。请注意，存在不同根的标签不平衡的问题，因此在本论文的模型中会通过调整正负权重的平衡以获得更好的结果。

规则集学习：

构建强大分类器的一个挑战是特征交互，当一些特征的值相互影响时，就会发生这种交互。特征交互的存在使得输出不能表示为单个特征效果的总和。决策规则，由一组“如果……那么……”组成逻辑规则，可以自然处理特征交互。具体来说就是规则是子句的逻辑连接（特征和阈值的比较），它在目标和特征之间建立逻辑关系，并且能够对特征和目标之间的非线性交互进行建模。规则集的另一个重要特性是其可解释性。规则的逻辑结构使其易于解释。规则的可解释性使得我们能够理解特征和目标之间的关系，并有助于检测导入特征。Skope规则使用树模型生成候选规则。它们构建了许多决策树，并将从根节点到内部节点或叶节点的路径作为候选规则。然后根据一些预定义的标准（如精确度和召回率）过滤这些候选对象。只有那些精确性和召回率高于阈值的人才会被保留下来。最后，采用相似性过滤方法选择具有足够多样性的规则。在解决方案中，应用Skope规则来学习每个根本原因的潜在规则，并删除预测样本不属于任何原因的规则。

预测归因模型：

当节点之间的相互依赖关系可用时，可以估计特征的重要性。特征重要性衡量向因果图中添加特定特征的边际收益。为此，论文生成了一个新的特征，用于衡量每个样本的特征重要性，并将它们集成到我们的模型中。特征重要性估计基于Shapley值。给定特征S的集合，internal和特征0的关系f，让是x的子集（只包含T中的特征）特征i的性状值为：



然而，直接计算Shapley值会带来两个困难。首先，当所有特征都准备好时，函数f只会生成一个输出，而不能仅估计给定部分特征的输出。其次，计算Shapley值非常耗时，因为它需要计算所有可能订单的边际收益。

为了解决这些问题，论文使用了一些近似，最终得到如下结果：

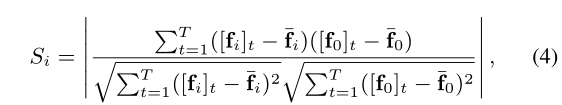


这样的近似效果很好，特别是对于稀疏因果图。在实验中，通过训练XGboost模型来估计内部节点和特征0节点之间的关系函数f。在估计特征重要性之后，通过简单地将其与预定义阈值进行比较来确定根本原因。那些重要性高于阈值的严重原因被确定为真正的根本原因。

图形算法：

利用所提供的因果图，对根本原因进行排序和定位：

第一个动机是，根本原因旁边的特性应该在相似性度量中显示出与目标特性0的高度相关性。由于特征0是关心的目标变量，且特征/KPI的值随时间变化并相互影响，因此将Pearson相关性的绝对值计算为特征i和特征0之间的相似性得分i，如下所示：



在计算皮尔逊相关性之前，对所有特征的缺失数据进行线性插值。皮尔逊相关性衡量两个特征（时间序列）如何随时间变化，并表示从0（不相关）到1（完全正或者负相关）的关系。此基于相关性的相似度评分将特征的相关性标记为目标特征0。由于相关性并不总是意味着因果关系，使用相似性得分可能会导致误报。相反，将相似度得分和因果关系图作为第二个动机来提高绩效。论文采用流行的图算法个性化PageRank来利用因果图。其主要思想是根据相似度得分对因果图进行随机遍历。具体来说，从特征0开始，通过随机选取因果图中的相邻特征，按顺序选择特征。拾取概率与边缘权重成正比，边缘权重由归一化相似性得分计算，其中如果两个特征fi和fj连接，则为1，否则为0。最终的根是基于这样的假设，即对根本原因旁边的功能的访问越多，就越可能是功能0的真正的根本原因。

**3. EXPERIMENTS AND DISCUSSION：**

论文在这一部分主要讨论了算法在竞赛中的表现情况。

**3.1. Datasets and Evaluation Metrics（数据集和评估指标）**

主要介绍了数据集的选取和评估的指标：竞赛提供的数据集包括一个固定因果关系图和特征数据集。在样本中，只有约45%被标记为根本原因故障，其他没有标记，这表明标签少并且不全面。对于评估指标，论文采用了竞赛所提供的标准化最终分数，即每个真阳性增加1分，而每个假阳性减少1分。最终分数根据测试样本的数量进行归一化，即最终最高的分数为1。

**3.2. Implementation and configuration（实施和配置）**

利用算法从原始数据中生成各种特征。然而，由于训练样本数量有限，仅针对所有这些特征训练模型就会导致模型过拟合，因此需要仔细地进行特征工程和选择。开始时尝试从训练数据中使用类标签集{root1，root2，root3，root2&root3}建立多类分类模型，然而，在这样的设置下，很难获得0.7分。

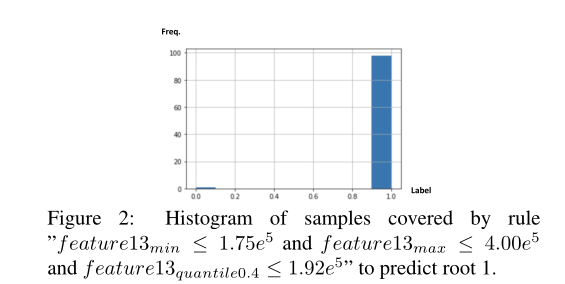
这种模式有几个缺点。首先，所有功能都在根本原因1、根本原因2和根本原因3之间简单共享。考虑到根本原因1时，无需添加特征20s或特征X、特征Y。其次，对于标签集{root1，root2，root3，root2&root3}，隐含的意思就是根本原因1和根本原因2不会同时出现，根本原因1&根本原因3或根本原因1&根本原因2&根本原因3也不会同时出现。但是事实上，这种假设可能不正确，或者说不具有普适性。

基于以上观察和实验结果，论文分别训练了三个二元分类模型：根本原因1、根本原因2和根本原因3。基于特征工程和数据扩充的衍生数据，对于根本原因1模型，主要使用来自特征0、13、15的信息以及它们之间的相互作用。当为根本原因1设置的特征从{0，13，15，19，20，X，Y}更改为{0，13，15}时，测试分数从0.825增加到0.837，验证了三种二元分类模型的有效性。对于根本原因2的模型，捕获有用信息的主要使用特征来自特征19和20。对于根本原因3的模型，使用生成的特征来捕获特征X和Y中的信息，因为特征Y包含重要信息来区分根本原因3和根本原因2。

根据因果图，可以看出不同的根本原因与不同的特征相关。为提高性能，论文采用集成建模，首先分别为root1、root2和root3训练具有上述不同特征的XGBoost模型，然后用规则集模型、属性模型和图模型进一步增强结果。

**3.3. Model Interpretability（对结果的解释）**

图2显示了由预测根本原因1生成的可解释规则之一覆盖的样本直方图，该规则的准确性接近于1。值得注意的是，标题中描述的布尔规则意味着特征13的较低值更有可能与根本原因1相关联。因果图证实，根本原因1通常与功能13和15相关的资源不足有关，这也与直觉相符。



**3.4. Performance Comparison and Ablation Studies**

论文首先将1407个标记样本分为大小分别为942和465的训练集和验证集。

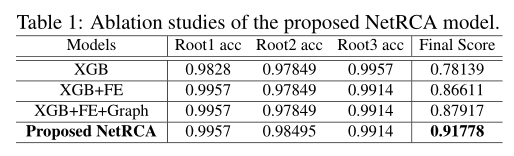


表1中为四种模型的比较，从第一行至第四行分别是：基本的XGBoost，在之前的基础上加上生成的特征（XGB+FE），在之前的基础上图算法（XGB+FE+Graph）,在之前的基础上加上数据扩充，规则集学习，属性模型（NetRCA）。表格的前3列表示模型在属于每个根本原因（1、2和3）的验证集上的性能准确性，第4列表示测试数据生成解决方案的分数。

分析表格可以得到一些结论：

1）所有模型，即使是基本的XGB模型，都可以在训练数据中得到很好的准确性。然而，提交的分数表明，培训和测试数据的分布之间存在一定的差距。三个模型显示出不同程度的过拟合，NetRCA算法可以防止过拟合并提供更稳健的解决方案。

2）XGB+FE模型优于基本XGB。通过深入业务背景，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信息，能够获得完整的视角并发现一些潜在的规则。

3） XGB+FE+Graph优于XGB+FE，其中，训练集没有显著的改进，但结合图模型可以将最终提交分数提高1%以上。原因可能在于图形模型可以更好地捕捉这些特征之间的因果关系。

4） 由于属于根本原因2的样本非常有限，因此特征工程和图模型似乎对训练集上根本原因2的准确性没有影响。但NetRCA显示，由于数据增加，根本原因2的准确性有了显著提高，这可以解决数据不平衡的问题。此外，识别根本原因2的另一个挑战在于根本原因2和其他根本原因的并发性。这样，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，进一步提高最终得分。

**4. CONCLUSIONS：**

这部分是对论文的总结，即本论文提出了一种新的网络故障根源定位算法NetRCA。除特征工程之外，该算法还采用了数据扩充来生成新的训练数据，以克服标记样本的不足。此外还设计了一种集成方法，有效地结合不同的模型，对网络故障进行准确可靠的因果推理。

**5. REFERENCES**

该部分列出了论文的参考文献，本篇论文共有23篇参考文献。论文中引用的算法等都内容在这些文献中都有记录，并且在本文中引用的部分也进行了标记。

**四、总结与体会**

网络故障根因定位挑战赛在上学期的课堂中我听老师提起过，在刚开始做本次论文阅读作业时，我本想选择我们北京交通大学的参赛论文进行研读，但是由于链接我没能打开，所以我选择了本篇论文进行研读。

本篇论文的结构比较清晰，运用课堂资料中老师教给我们的一些方法，比如仔细阅读主题句，先对文章整体或者对应的段落有一个大致的了解，然后再仔细阅读细节部分，比如算法的阐述，或者对实验结果的分析等等，同时注意论文中出现的图，比如本篇论文中的图1是算法的结构图，上面标明了许多关键的字词，同时该图片其实也是论文第二部分的结构，论文的第二部分围绕着这幅图介绍算法。

通过本专题研讨课，我已经阅读了较多篇英文文献，对文献的阅读有了一定的了解。我记得我第一次阅读英文文献的时候非常紧张，读了好久也翻译不出来，并且不能理清论文的行文结构，再加上论文本身具有专业性，这让我没有头绪。而当阅读本篇论文的时候，我先了解一下文章的整体结构，再从细节入手进行研究，与之前比较有了一定的进步。非常感谢课程提供给我们的这些锻炼的机会，从中我提高了自己的综合能力和专业素养。