一、文献信息：

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

论文题目：NETRCA: An effective network fault cause localization algorithm

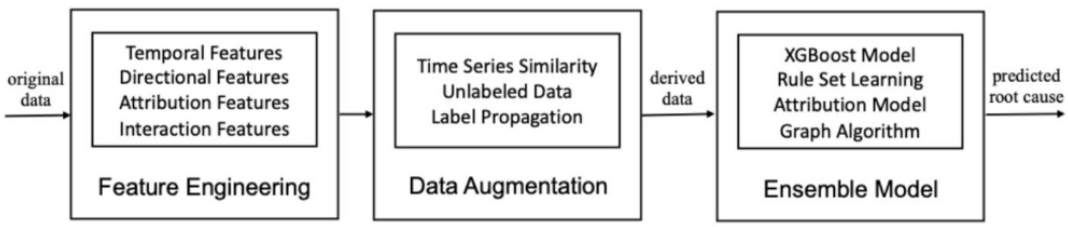
发表途径：ICASSP 2022 5G 网络故障根由定位挑战赛;本篇论文为阿里达摩院第一名论文。

二、问题意义：

研究问题： 5G 无线网络故障根因定位，快速准确地找出网络故障的原因

研究背景：现实中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，快熟准确地定位故障的根因，是网络运维面临的巨大挑战。此篇论文即在这个特定背景下，提出的新的方法来解决网络故障定位问题。

主要工作：在本文中，作者提出了一种新的 NetRCA 算法来解决这个问题。首先，通过考虑时间、方向性、归因和交互特征，从原始数据中提取有效的衍生特征。其次，采用多元时间序列相似度和标签传播，从标记和未标记数据中生成新的训练数据，以克服了标记样本缺乏的问题。第三，设计了一个结合了 XGBoost、规则集学习、归因模型和图算法的集成模型，以充分利用所有的数据信息，提高性能。最后，对 ICASSP2022AIOps 挑战的真实数据集进行了实验和分析，证明了该方法的优越性和有效性。



NetRCA框架

论文中提出的 NetRCA 框架主要包括了（Feature Engineering）特征工程模块、（Data Augmentation）数据增强模块和集成模型（Ensemble Model）的方式。

三、思路方法：

基于通信网络 ICASSP2022AIOps 挑战的数据集，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。NetRCA 由三个主要组件组成，包括特征工程、数据增强和模型集成。在特征工程中，为时间序列数据和与无线方向相关的特征设计了特征。然后，又由于标记数据有限，而存在大量的未标记数据，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。

具体实现步骤：

1. Feature Engineering（特征工程）

基于从每个样本中提取的特征来训练模型。生成的特征大致可分为四类：时间特征、方向相关特征、归因特征和交互特征。

捕获特征之间的相互关系的特征，得到统计特征，然后从每个时间切片样本的距离分布中总结为模型训练。生成新的特征作为每个特征相对预测特征 0 的重要性得分估计。然后生成 X 和 Y 的二阶交互特征。计算 X 与 Y 的比值。最后计算这些比率的统计数据，作为对时间特征所做的事情。

2. Data Augmentation（数据增强）

从与标记数据相似性较高的未标记数据中选择样本，并根据与训练集相似的训练样本的真实根源对训练集进行标记。对每种类型的根本原因分别执行此过程，以提高计算效率。将所有训练样本的时间戳对齐，并将它们的真实标签增强为所有根原因标签的并集。

3. Ensemble Model（集成模型）

NetRCA 采用集成模型对根本原因进行预测，采用 XGBoost 获得初始结果，然后结合规则集学习、归因模型和图解算法对结果进行细化，得到最终结果。

四、实验结论：

作者根据自己的模型及算法，总结讨论了 netRCA 在题目所给数据集上的性能表现。

1. Datasets and Evaluation Metrics 数据集和评估指标

数据集包括一个固定的因果关系图和特征数据集，其中包含 2984 个样本和 23 个可观测变量。在 2984 个样本中，只有约 45%的样本被标记为根本原因故障，而其他样本仍未被标记，说明标签稀缺且不全面。对于评价指标，我们采用挑战提供的归一化最终分数，每个真阳性根增加 1 分分数，每个假阳性根扣除 1 分分数。最终的分数由测试样本的数量归一化，因此最终的最高分数为 1。

2. Implementation and configuration 实施和配置

尝试了使用来自训练数据的类标签集{根 1、根 2、根 3、根 2 和根 3}的多类分类模型，然而，在这样的设置下，得分是 0.7+。这种模型有几个缺点。首先，所有的特性都只是在根本原因 1、根本原因 2 和根本原因 3 之间共享的。当考虑根本原因 1 时，不需要添加特征 20s 或特征 X，特征 Y。其次，标签设置为{根 1、根 2、根 3、根 2 和根 3}，一个隐式的语句是根原因 1 和根本原因 2 不会同时出现，根本原因 1 和根原因 3 或根本原因 1 和根本原因 2 和根本原因 3。在现实中，这种假设是有限的，而不是普遍的。

为了提高整体性能，采用集成建模，首先对具有不同特征的 root1、boost 模型进行训练，然后通过规则集模型、归因模型和图解模型进一步增强结果。

3. Model Interpretability 模型可解释性

此部分演示了可解释性在帮助诊断模型性能和提高人类信任度方面的有效性。

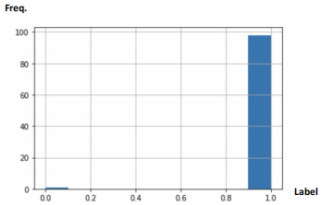


Figure 2: Histogram of samples covered by rule "$feature13_{min} \leq 1.75e^5$ and $feature13_{max} \leq 4.00e^5$ and $feature13_{quantile0.4} \leq 1.92e^5$" to predict root 1.

上图是被预测根本原因 1 的可解释规则所覆盖的样本的直方图。很明显，该规则的准确性接近于一。更重要的是，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 有关。因果图证实了根本原因 1 通常与与特征 13 和特征 15 相关的资源不足有关，这与规则背后的直觉相匹配。

4. Performance Comparison and Ablation Studies 性能比较和消融术研究

Table 1: Ablation studies of the proposed NetRCA model.

Models	Root1 acc	Root2 acc	Root3 acc	Final Score
XGB	0.9828	0.97849	0.9957	0.78139
XGB+FE	0.9957	0.97849	0.9914	0.86611
XGB+FE+Graph	0.9957	0.97849	0.9914	0.87917
Proposed NetRCA	0.9957	0.98495	0.9914	**0.91778**

表的前 3 列表示模型在属于每个根本原因（根本原因 1、2 和 3）的验证集上的性能的准确性，而第四列表示在测试数据上生成的解决方案提交的分数。

可以看到：1)三种消融模型都表现出不同程度的过拟合，而 **NetRCA 算法可以防止过拟合，**并给出一个更鲁棒的解决方案。2)**XGB+FE 模型在训练和验证集上都显著优于基本的 XGB。**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，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**结合图解模型可以使最终的提交分数提高 1%以上**。原因可能在于，图解模型可以帮助更好地捕捉这些特征之间的因果关系。4)特征工程和图模型似乎对根本原因 2 对训练集的准确性都没有影响，因为属于根本原因 2 的样本非常有限。但是**最终的 NetRCA 显示，由于数据的增加，根本原因 2 的准确性显著提高。**此外，识别根本原因 2 的另一个挑战是根本原因 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，并进一步提高最终得分。

五、启发思考：

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，算法采用数据增强来生成新的训练数据，以克服标记样本的缺乏。此外，作者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。

其实单论每个小的点来说，都是我们熟悉的或者说不是很陌生不可以接触到的东西，但是作者能够很好地将它们结合到一起，并构造出最终的模型进行问题的求解。也就是说，在我们的平常学习和各种方面，要学会前后知识的融会贯通，举一反三，也做到各种知识的对比收集，在面对具体的问题时，能够要想到什么是适合的什么是不适合的，才是正确的学习方法和学习态度。