信息网络综合专题研究

应用层阅读报告

课 程 名 称 ：	信息网络综合专题-应用层
学 生 姓 名 ：	王海旭
学　　号 ：	19211148
班　　级 ：	通信 1905
任 课 教 师 ：	陈一帅、郭宇春
日　　期 ：	2022.5.7

目录

一、 概述：

无线网络中故障根因定位对网络运行和维护至关重要，但在实现上存在一定的困难。受限于复杂的无线通信环境和网络部署结构，加之网络故障样本数少、不同的场景下故障表征差异性大等问题，这一目标并不容易实现。这篇论文提出一种新的算法 NetRCA，并且通过实验来证明了这个算法的优越性与有效性。

这篇论文整体结构为提出问题、提出方法、方法实现、方法验证，科学有效地综合多种已有技术，达到“1+1>2”的效果。关键词为：根本原因分析、数据扩充、时间序列、集成模型、无线网络。

二、 导言：

在处理这一问题上，已经有很多人尝试过不同的方法，原文中举证多个论文中提出的贝叶斯网络、无监督算法、时序算法、分层贝叶斯网络等多种技术，然而由于技术限制，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辄止，不能很好地解决问题。

作者将问题的难点归结为三类，分别是：①网络深度的增加可能会使从源节点到根节点的路径中传播错误，从而使根本原因定位变得困难；②缺乏足够的已知标签；③与每个网络节点相关联的时间序列数据是多元的，通常处于复杂模式，具有相互依赖性和噪声，难以提取节点关系。

针对上述问题，文中提出的 NetRCA 算法设计了三个主要组件，分别为特征工程、数据扩充、模型集成，生成了多个标记数据，利用规则集学习、归因模型、图形算法、因果关系图提高性能，使得算法模型最后的输出具有较高的预测精度。

三、 算法框架：

1. 整体框架

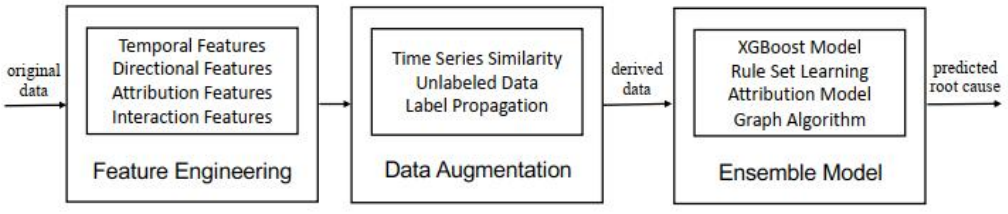

图 1　提出的 NetRCA 算法的框架

原始输入数据依次通过特征工程、数据扩充和模型集成，输出推断的根因。

2. 特征工程

这里先解释时间戳的概念：时间戳是使用数字签名技术产生的数据，签名的对象包括了原始文件信息、签名参数、签名时间等信息。

由于每个样本之中包含的时间戳数量不同，直接使用时间戳用于模型训练，最终的模型可能倾向于具有更多时间戳索引的样本，这就造成了预测的偏差。因此，为了训练模型的准确性，需要使用从每个样本中提取出来的特征来训练模型。这些特征可以分为以下四类，分别是时间特征、方向相关特征、归因特征、交互特征。

时间特征：模型中涉及的时间特征都是基于数据统计的，每个时间戳中包含的数据被认定为相互独立，提取出平均值、最小值、最大值、中位数、十分位数、偏度。同时还包括一些反应时间序列形状的函数，反映出峰值数目和平均变化值。

方向相关特征：波束形成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，这对于根因定位也很重要。假设有每个节点从 0 到 31 的索引，映射到一个 4×8 的位置矩阵，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。最后从每个时间片样本的距离分布中总结出统计特征。

归因特征：根据因果图，推导出除特征 0 之外的所有节点的属性特征，这些根本原因最终导致功能 0 的值降低。真正的根因及后续影响将比其他因素对功能 0 的当前值影响更强。因此生成一个新特征，作为预测特征 0 上每个特征的重要性得分的估计值。

交互特征：生成 X 和 Y 的二阶交互特征。当特征 X 等于特征 Y 与一些未知因素的比值时，生成特征 X/Y 来衡量这些未知因素的影响。首先根据问题描述将 X 和 Y 中的特征分组成对。对于每一对计算 X/Y。最后计算这些比率的统计数据。

3. 数据扩充

多元时间序列相似性：为了度量不同长度的多元时间序列之间的相似性，采用 Eros 算法来计算相似性。Eros 通过使用主成分并基于特征向量计算相似度来扩展 Frobenius 范数。其中 Frobenius 范数的定义为矩阵的每个元素的平方和的

开方，数学表达为$\|A\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|a_{ij}\right|^2}$。

形式上，假设 A 和 B 分别是大小为$m_A\times n$和$m_B\times n$的两个矩阵。通过对 A 和 B 的协方差矩阵分别应用奇异值分解，设$V_A=[a_1,\cdots,a_n]\in R^{n\times n}$和$V_B=[b_1,\cdots,b_n]\in R^{n\times n}$为两个右特征向量矩阵。则 A 和 B 的相似性是

$$Eros(A,B,w)=\sum_{i=1}^{n}w_i|<a_i,b_i>|,(1)$$

其中，$w=[w_1,\cdots,w_n]$是基于特征值的权重向量，满足$\sum_{i=1}^{n}w_i=1$。

数据/标签扩充：标签数据通常是有限的，因此数据扩充是必要的。超过一半的训练数据是未标记的，而删除这些数据又会丢失很多有价值的信息。使用 Eros 算法测量任意两个训练样本之间的相似性。在未标记数据中，选择那些与标记数据高度相似的样本来丰富训练集，并根据这些训练样本的真正根因进行标记。该过程分别针对每种类型的根因执行，以提高计算效率。

另一个重要的扩充是传播共享相似时间戳的训练样本的根因标签。这改进了对测试数据集的多个根因的预测。若几乎所有被标记为根因 1 的训练样本都与其他根因无关，则可以假设根因 1 独立于其他根因发生。然而，仔细观察所有训练样本的时间戳和标签，在根因 1 与根因 2 和 3 同时发生的情况下，存在大量一分钟的时间间隔。因此将所有训练样本按其时间戳对齐，并将其真实标签扩充为所有根本原因标签的并集。

4. 模型集成

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。详情如下：

通过 XGBoost 进行根本原因分类：将找到正确的根因视为一个分类问题。具体来说，采用 XGBoost 作为基础模型。由于存在不同根的标签不平衡的问题，因此在模型中调整了正权重和负权重的平衡，以获得更好的结果。

此处对 XGBoost 进行粗略介绍，其思想为：总共构建 T 颗树。当构建到第 t

颗树的时候，需要对前 t-1 颗树对训练样本分类回归产生的残差进行拟合。每次拟合产生新的树的时候，遍历所有可能的树，并选择使得目标函数值最小的树。但是这样在实践中难以实现，因此需要将步骤进行分解，在构造新的树的时候，每次只产生一个分支，并选择最好的那个分支。如果产生分支的目标函数值比不产生的时候大或者改进效果不明显，那么就放弃产生分支。可以并行化处理。

规则集学习：构建分类器的一个难点是特征交互，当某些特征的值相互影响时，就会出现这种问题。这使得输出无法表示为单个特征效果之和。规则集的另一个重要属性是可解释性，规则的逻辑结构使其易于解释。规则的可解释性有助于对导入的特征进行检测。

本文中使用 Skope 规则，构建了大量决策树，将从根节点到内部节点的路径作为候选规则。随后根据一些预先制定的标准过滤这些对象。只有那些高于标准阈值的对象会被保留下来。最后，应用相似性过滤来选择具有足够多样性的规则。

预测归因模型：当节点之间的相互依赖关系可用时，可以用于估计特征的重要性。特征重要性衡量向因果图中添加特定特征的收益。上游节点中的异常数据可能会导致功能发生很大的变化，这可以帮助确定根本原因。

在实现过程中，生成了一个新的特征，用于测量每个样本的特征重要性，并将它们集成到模型中。这个特征的重要性估计基于 Shapley 值。给定一组特征 S，内部和特征 0 之间的关系 f，设 x_T 为仅包含 T 中特征的 x 的子集，特征 i 的 Shapley 值 $\phi(i)$ 为

$$\phi(i)=\sum_{T\subseteq S\setminus\{i\}}\frac{|T|!(p-|T|-1)!}{p!}(f(x_{T\cup\{i\}})-f(x_T)),(2)$$

此时对于 Shapley 的含义未进行介绍，查阅资料后，做出较为易懂的解释。

假设这样一个场景：一群拥有不同技能的参与者为了集体奖励而相互合作。那么，如何在小组中公平分配奖励？

在这一场合中，存在一组 N 个参与者。有一个函数 v，给出了这些参与者的任何子集的值，即 S 是 N 的子集，然后 v(S)给出了该子集的值。对于一个联合博弈(N,v)，可以使用这个方程来计算参与者 i 的贡献，即 Shapley 值。

在合作中增加参与者 i 的边际价值。对于任何给定的子集，比较它的值和当包括参与者 i 的时候它的值。这样做得到了将参与者 i 添加到该子集的边际值。

在本文中，测量了以不同顺序添加特征 i 的平均边际增益。然而，直接计算 Shapley 值存在两个困难。首先，函数 f 只在所有特征都准备好时才有输出，且不能仅估计给定部分特征的输出。其次，计算 Shapley 值需要计算所有可能特征的边际收益，这会让耗时增加。

为了解决这两个问题，作者使用 $f(x_T, \bar{x}_{S\setminus T})$ 来近似 $f(x_T)$，其中 $\bar{x}_i$ 表示特征 i 的平均值。换言之，$f(x_T)$ 近似为输入 $[x_T, \bar{x}_{S\setminus T}]$ 时 f 的输出，此时保持 T 中的特征不变，并将剩余特征设置为其平均值。这是用于计算 Shapley 值的常用策略。为了克服第二个困难，将 Shapley 值近似为从 S 中移除 i 时 f 的值减少量，即

$$\phi(i) \approx \left| f(x_S) - f([x_{S\setminus\{i\}}, \bar{x}_i]) \right|, (3)$$

对于稀疏的因果图来说，这样的近似效果很好。在文中后续实验里，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，简单地将其与预先设置阈值进行比较来确定根因。重要性高于阈值的被确定为真正的根因。

图算法：为了进一步利用给定的因果图，原文设计了一种基于单变量时间序列相似性的特殊算法，来对真正的根因进行排序和定位。第一个动机是，根因旁边的特征应该在相似性度量中显示出与目标特征 0 的高度相关性。由于特征 0 是操作者关心的目标变量，且特征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 Si，如下所示：

$$S_i = \left| \frac{\sum_{t=1}^{T} ([f_i]_t - \overline{f}_i)([f_0]_t - \overline{f_0})}{\sqrt{\sum_{t=1}^{T} ([f_i]_t - \overline{f}_i)^2} \sqrt{\sum_{t=1}^{T} ([f_0]_t - \overline{f_0})^2}} \right|, (4)$$

其中 fi 是特征 i 的单变量时间序列数据，fi 表示其平均值。在计算 Pearson 相关性之前，需要对所有特征的缺失数据进行线性插值。Pearson 相关性衡量两个特征如何随时间变化，并表示从不相关到完全正/负相关的关系。这种基于相关性的相似性得分 Si 表示特征 i 与目标特征 0 的相关性。由于相关性并不总是意味着因果关系，使用相似性评分可能会导致误报。相反，相似性评分和因果关系图都是提高效果的第二个动机。于是文章又采用图算法个性化 PageRank 来利用因果图。主要思想是根据相似性分数在因果图上进行随机游走。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，边缘权重通过标准化相似性得分计算为 $w_{ij} = A_{ij}S_j / \sum_j A_{ij}S_J$，其中，如果两个特征 fi 和 fj 相关，则 Aij 为 1，否则为 0。这表明对根因旁边的功能的访问越多，根因就越有可能是功能 0 的真正根。

四、 实验讨论：

在这一部分中，文章总结并讨论了在给定数据集中 NetRCA 算法的性能。

1. 数据集和评估指标

给定的数据集包括一个固定因果关系图和特征数据集，其中包含样本和可观察变量。在样本中，只有约 45%的样本被标记为根因故障，而其他样本则没有标记，这说明标签稀少且不全面。

对于评估指标，文章采用挑战提供的标准化最终分数，即每个真阳性根因增加 1 分，而每个假阳性根因减少 1 分。最终分数通过测试样本的数量标准化，因此最高最终分数为 1。

2. 实现与配置

上文已经介绍过，算法从原始数据生成各种特征。然而，由于训练样本数量有限，仅针对这些特征来训练模型，会导致模型过度拟合。这就需要在进行特征工程和选择时做出更多处理。

原文首先尝试了基于列车数据的多类分类模型，其中类标签集为{root1、root2、root3、root2&root3}。然而，在这种模型下很难获得 0.7 以上的分数。这种模型有几个缺点。首先，所有功能都在根因 1、根因 2 和根因 3 之间共享。在考虑根因 1 时，无需添加特征 20s 或特征 X、特征 Y。其次，将标签设置为{root1、root2、root3、root2&root3}，隐含声明是根因 1 和根因 2 不会同时出现，在现实中，这种假设不具有一般性。

基于上述实验结果，作者分别训练了三个二元分类模型，分别用于根因 1、根因 2 和根因 3。基于特征工程和数据扩充的衍生数据，对于根因 1 模型，选取了来自几个特征的信息以及他们之间的交互作用。此时在相同参数下，测试分数从 0.825 增加到 0.837，这证明了三个二元分类模型的有效性。对于根因 2、根因 3 的模型同理，使用生成的特征捕获特定特征内的信息。

根据给定的因果图，可以看出不同的根因与不同的特征有关。为了提高整体

性能，论文采用集成建模。首先训练具有上述不同特征的 root1、root2 和 root3 的 XGBoost 模型，然后通过规则集模型、属性模型和图模型进一步增强结果。

3. 模式可解释性

接下来文章展示了模型可解释性在帮助判断模型性能和提高可信度方面的有效性。图 2 是为预测根因 1 而生成的可解释规则之一所对应的样本直方图。这条规则的准确性接近 1。通过应用 NetRCA 提出的可解释模型，可以更深入地理解模型是如何进行预测的，并纠正为错误原因预测正确答案的问题。

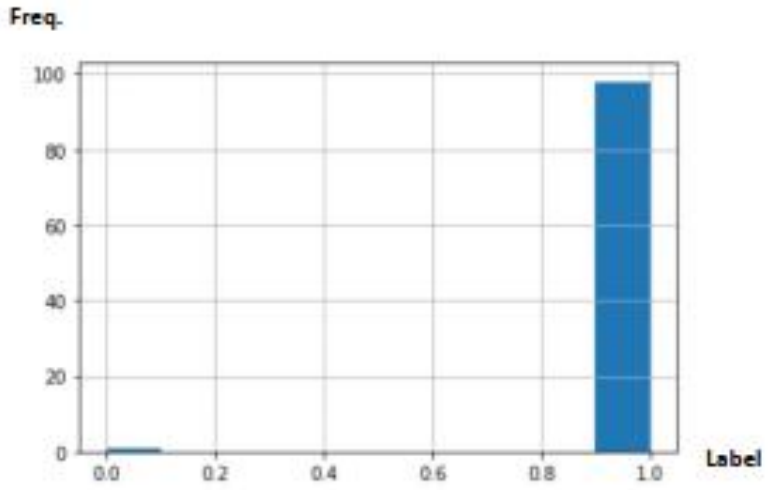

图 2　基于规则

“ $feature13_{\min} \le 1.75e^5$、$feature13_{\max} \le 4.00e^5$、$feature13_{quantile0.4} \le 1.92e^5$ ”对根因 1 进行预测的直方图

4. 性能比较与消融实验

消融实验通常用于神经网络，可以理解为设置对照组。通过去除某个模块的作用，来证明该模块的必要性，如果消融实验后得到的结果不好或者性能大幅下降，说明该模块起到了作用。

消融实验过程中，原文将 1407 个标记样本分为训练集和验证集，大小分别为 942 和 465。在表 1 中比较了没有任何额外特征的基准线 XGBoost 模型、具有前文所述特征工程中生成的特征的 XGBoost 模型、XGB+FE 和图形算法的组合，以及将 XGB+FE+图形与数据扩充、规则集学习和属性模型相结合的 NetRCA 算法。前三列表示模型在属于每个根因的验证集上的性能准确性，第四列表示测试数据生成解决方案的得分。

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 1　NetRCA 模型的消融实验

从表 1 中的结果可以得到几个结论：

1）所有模型都可以在训练数据中获得极好的准确性。但是最终得分表明，训练数据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，而文章提出的 NetRCA 算法可以防止过拟合，提供更稳健的解决方案。

2）经过特征工程的处理，XGB+FE 模型在训练和测试数据集上都显著优于基本 XGB。从时态特征、方向相关特征、属性特征和交互特征中提取有效信息，操作者能够找到真正根因并发现部分潜在规则。

3）尽管训练数据集没有进行显著修改，但结合图模型可以将最终得分提高 1%以上。原因在于图形模型可以更好地捕捉特征之间的因果关系。

4）特征工程和图形模型似乎都不会影响测试数据集中根因 2 的准确性。这是因为属于根因 2 的样本数量有限。但最终结果显示，随着数据增加，根因 2 的准确性显著提高，意味着这可以解决数据数量不平衡的问题。

五、 总结：

这篇论文提出了一种崭新的网络故障根源定位算法 NetRCA。除了精心设计的特征工程之外，还采用数据扩充来生成新的训练数据，从而克服样本的不足。此外，论文还提出了一种有效地结合不同模型的集成方法，综合各家之长，对网络故障进行准确可靠的因果推断。

在对这篇论文进行阅读研究的过程中，我经历了一个提出问题、寻找思路、想法验证、解决问题的过程。我直观感受到了机器学习中诸多算法、思想在工程之中的实际应用。

已知在无线网络中很难实现故障根因定位，已存方法又都存在一些弊端，因此论文中提出的 NetRCA 更像是集百家之长的一种算法。在找到大致方向后，开始对算法进行细节设计，主要包括特征工程、数据扩充和模型集成。其中细节包括但不限于 Eros 算法扩展范数、XGBoost 算法、Shapley 值计算、Pearson 相关

性等对我来说未知的知识。在阅读的过程中，我对上述算法知识进行了浅层次的学习与理解。

最后是设置对照组进行消融实验，这部分体现了整个实验过程的科学性与严谨性，最终以表格形式直观反映出多个模型的区别，以此为依据，论证 NetRCA 算法在准确度、稳健性上的优势。

感谢老师们的教学，老师严谨求实、勤于思考的治学态度令我受益匪浅。在这一模块中学到的 AI 相关知识，以及在阅读论文过程中总结的学习方法，都将为我未来的学习、工作提供帮助。