应用层阅读报告

通信 1905 19211145 苏小宁

一、文献信息

论文题目：ACCURATE INFERENCE OF UNSEEN COMBINATIONS OF MULTIPLE ROOTCAUSES WITH CLASSIFIER ENSEMBLE

发表途径：ICASSP 2022 5G 网络故障根因定位 挑战赛

作者：北京交通大学电子与信息工程学院网络智能实验室

发表时间：2022 年

二、问题概述

2.1 研究意义

网络故障的根因定位在实际网络运维中具有重要意义，当 5G 无线网络发生故障时，我们需要快速对故障进行根因定位，只有快速准确找出故障的根本原因，才能及时采取措施对网络进行修复。当下对于网络故障根因定位面临通信环境和网络结构复杂、网络故障样本数少、故障无规律发生等问题。这篇论文提出了一种新的根因定位方案，通过单因/多因判断和采用多个二元分类的方式简化了分析复杂度，在时间片样本上得到了良好的测试效果。

2.2 数据集说明（竞赛官网）

1、特征因果图

图中顶部的特征 0 为故障检测的目标变量，当特征 0 的值较低时我们需要根据特征的因果关系图（图 1）分析影响因素，推断出导致网络故障的根本原因。

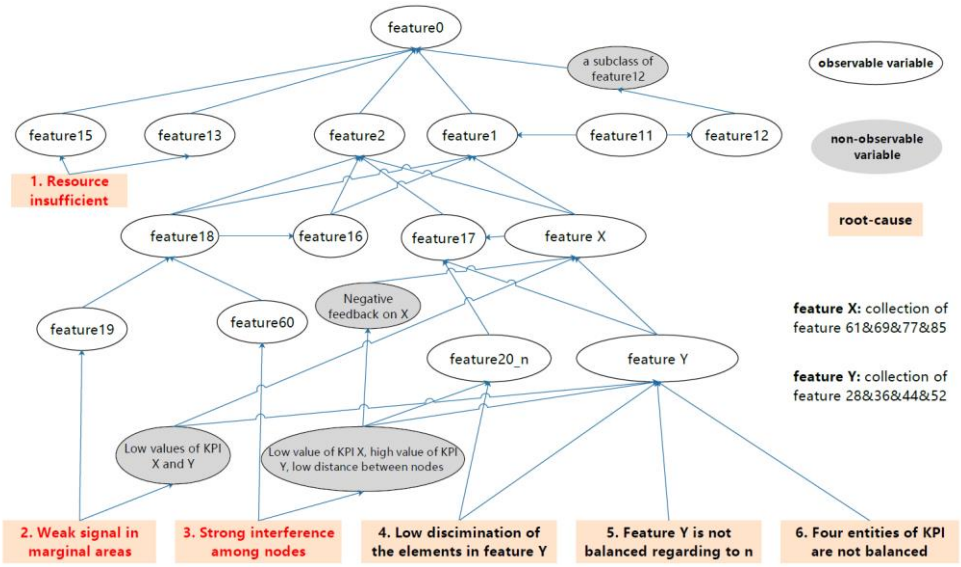

图 1 特征因果关系图

关系图说明：图中的特征因果关系来自于一个标准的通信协议，在不同的场景下具有通用性。椭圆形代表一个变量或一组变量，白色椭圆代表观察到的变量，灰色椭圆表示不可观察的变量，矩形框代表故障发生的几种根本原因。

2、实验数据集（时间片）

数据集共包 2984 个样本，每个样本都是从不同 5G 网络中测得的时间片。时间片的采样间隔为 1 秒，包括 23 个可观察变量（图 1 中的白色椭圆）的不同 KPI（关键性指标）信息。在数据集中，只有 45%的样本标记出了故障的根本原因，其余样本未被标记。

三、研究步骤

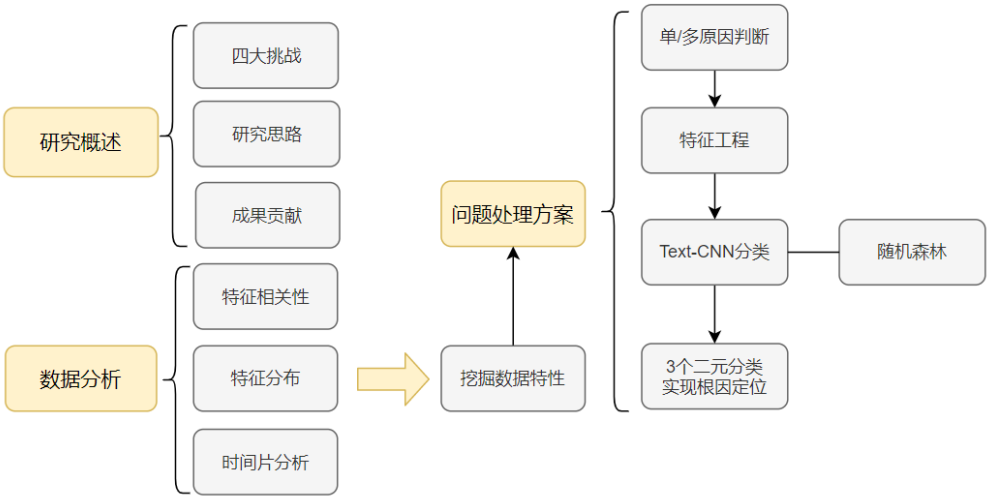

图 1 论文研究步骤图

论文分析思路概述：

论文整体上分为研究概述、数据分析、问题处理方案三个部分，首先作者分析无线网络根因定位的四大难点，给出了一个可行的问题处理思路。第二部分是数据分析，作者对原始数据进行了初步处理，观察了不同特征的相关性、不同根本原因下的特征分布、时间片样本特性，挖掘了一些样本数据的特性，得到了一些有助于后续问题处理的结论。在此基础上作者给出了根因定位的具体方案，首先进行单因/多因判断，进一步完成特征工程处理，最后通过 Text-CNN 网络分类。进一步作者通过比对多种机器学习模型的训练效果，给出了最终方案对根本原因 2 和 3 使用随机森林，对根本原因 1 使用 Text-CNN 分类，整体解决方案的得分为 0.93。

四、研究内容

4.1 研究难点与思路

首先作者总结了无线网络故障根因定位面临的四大挑战：

1.特征类型多样：数据集中有近 90 个特征，还包含一些不能直接处理的非数值特征。

2.时间片长度不同：样本时间片的长度范围在 1 到几百万，因此很难在时间维度上提取不

同根本原因的特征。

3. 故障可能由多个根本原因导致：当多个根本原因同时出现时，会影响到数据分布，从而增大了分类的难度。

4. 对于多种原因的组合故障缺少训练数据：在训练样本中只包含四个根本原因的组合（1，2，3，2&3），对于其他的根因组合如 1&2、1&3，则没有训练数据。

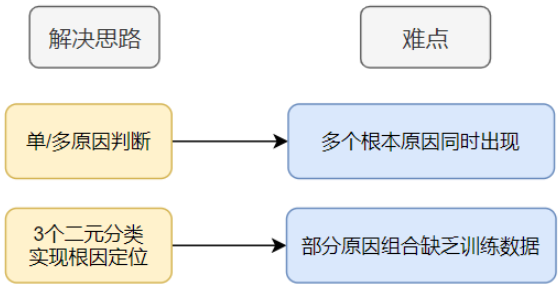

图 2 解决思路-难点

在描述四大难点之后，作者简要说明了分析流程。结合后续内容，我总结了一下作者的研究思路主要包括以下两个重要步骤。1、通过在数据处理前进行单因/多因判断，解决了多个根本原因同时出现的问题。2、通过 3 个二元分类（对于三种根本原因）解决了部分原因组合缺乏训练数据的难点。

4.2 数据分析

概述：在这一部分中作者对原始数据进行了初步分析，或者说是对特征进行了观察。作者分析了不同特征的相关性、不同根本原因下的特征分布、时间片样本特性，得到了一些有助于后续分类的结论（寻找关键特征的方法、判断是否为多种根本原因的方法）。

1、特征相关性分析

论文中给出了不同特征的相关性热度图，从图中可以看出特征 13 和 15 具有很高的相关性。进一步通过观察图一的特征关系图，可以发现特征 13 和特征 15 与根本原因 1 是直接相连的，因此论文中将它们定义为关键特征进行进一步分析。

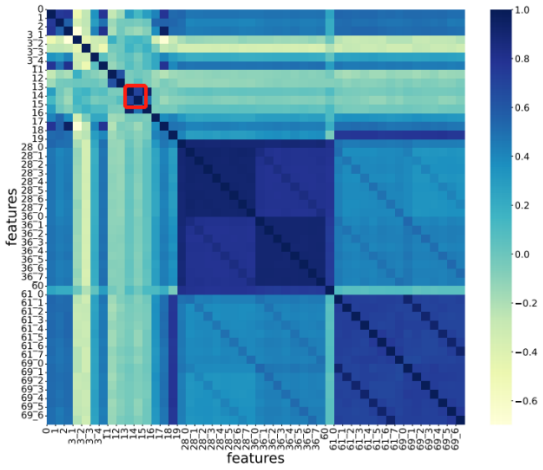

图 3 特征相关度热图

2、特征分布

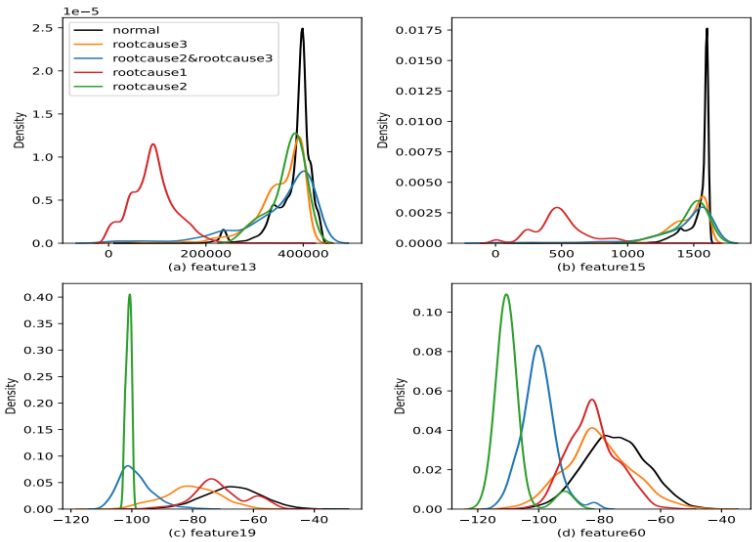

图 4 不同根本原因下的特征分布

论文中给出了四种不同特征（13、15、19、60）在不同根本原因下的分布图，并得出了以下两个重要结论：

1、特征 13 对于根本问题 1 的分布明显异于其他三种根本原因，因此可以认为特征 13 是根本原因 1 的关键特征。

2、特征 60 对于根本问题 2 和 3 的分布基本相近，因此我们不能将特征 60 作为根本原因 2、3 分类的依据。

3、时间片分析

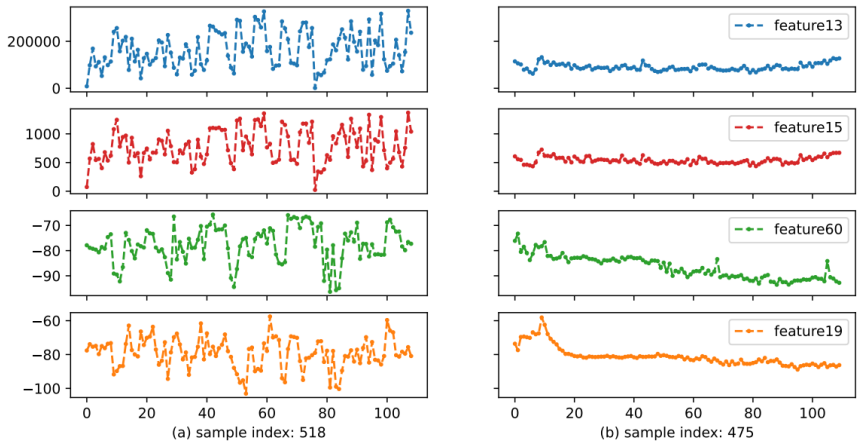

图 5 时间片-特征波动

上图展示了两个不同的样本中特征值随时间变化的图像，可以看出样本 518 的特征值波动性明显高于样本 475，因此作者推测样本 518 的故障可能包含多种根本原因。

4.3 分析方案

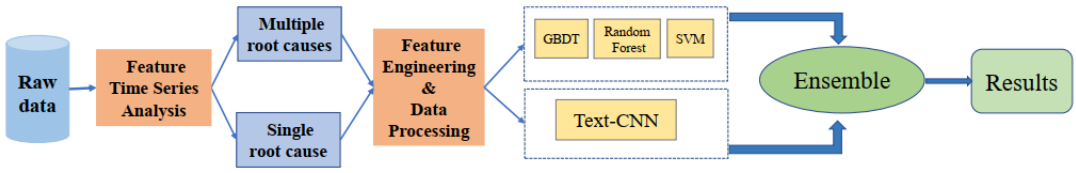

图 6 整体处理流程图

方案简述：上图展示了整体的论文分析流程，首先对特征进行时间序列分析，判断出导致故障的是单一原因还是多种原因。进一步进行特征工程和数据处理，对于处理后的数据作者采用了 Text-CNN、SVM、GBDT 四种方法进行分类，最终给出了一种高成功率的方案。

1、判断单一原因 or 多种原因

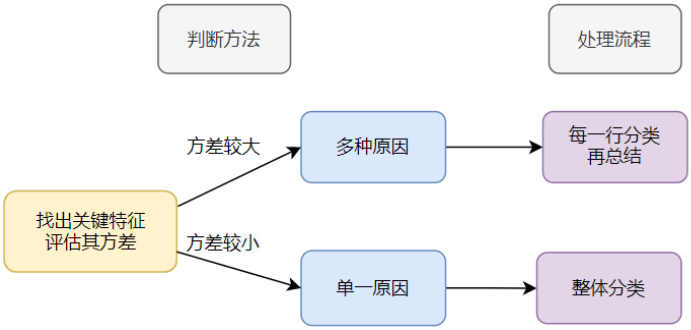

图 7 不同情况分类流程图

首先，根据特征因果关系图选取不同根本原因的关键特征（相关度高）。然后在样本中评估其方差。若方差较大，则推断故障是基于多种根本原因。反之，则判断为单一原因。对于多种根本原因的情况，作者采用每一行进行原因分类再总结的方案。对于单一原因，则直接进行整体分类。

2、特征工程（空间特征设计）

下图（a）展示了特征值的空间分布，图中包含 8 个方向的接收值（特征 20），数值为 0-31（整数），构成了一个 4*8 的矩阵。

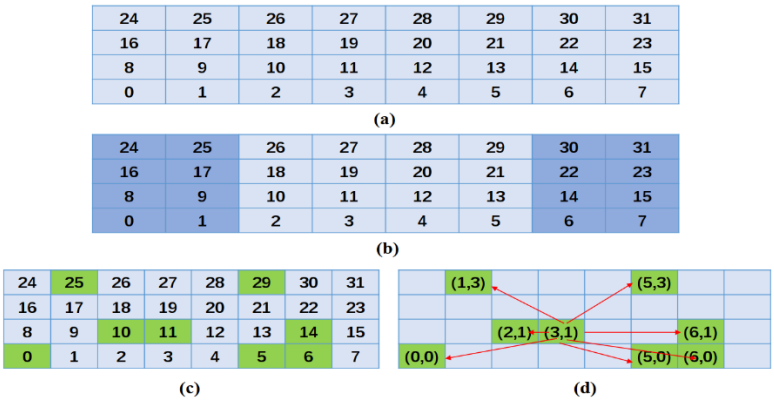

图 8 空间特征设计

在此基础上，作者设定了两个特性来实现特征刻画。

1）特征边缘（feature_edge）

根本原因 2 涉及到边缘区域的弱信号，因此作者设计了特征边缘这一指标来表征这一特性。作者给出了具体刻画方案：首先取矩阵中最左侧的两列和最右侧的两列作为边缘列，如图（b）。进一步在每个时间点统计 8 个方向特征值与边缘列相碰撞的次数（这里我的理解是统计数值相等的次数），将统计值作为特征边缘。

2）特征距离（feature_distance）

根本原因 3 涉及到节点间的强干扰，由于信号间的干扰强度可以用接收方向之间的距离来表征，作者设计了特征距离这一特性。首先将数值矩阵放入直角坐标中，用坐标值表示矩阵的行列位置，进一步可以通过计算每两个节点之间的欧氏距离得到特征距离。

3、基于 Text-CNN 进行原因分类

1）预处理

Text-CNN 网络要求输入数据的向量长度是统一的，作者首先对每个样本的特征值长度进行填充和裁剪，使其大小一致。由于 75%的实验样本长度在 30 左右，因此作者将 30 定位输入数据的长度。对于长度超过 30 的样本，进行数据剪切；对于长度小于 30 的样本，作者采用特征的平均值进行填充。

2）网络设计

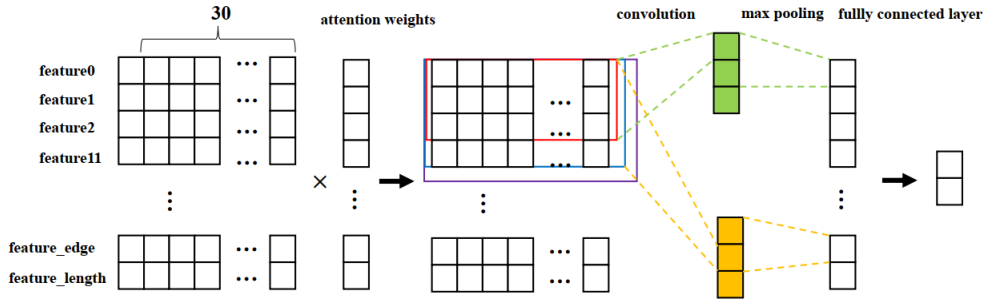

图 9 Text-CNN 网络模型

（1）注意力层（Attention layer）

由于不同的特征在原因分类过程中的具有不同的重要性（权重不同），因此作者引入了注意力层来动态地计算特征权重（通过模型训练的方式）。计算方式如下所示：其中 W 为可训练的参数，e、h 分别为注意力层的输入与输出。

$$e_n = W_n^T \cdot f_n$$

$$\alpha_n = \frac{exp(e_n)}{\sum_{j=0}^{nums} exp(e_j)}$$

$$h_n = \sum_{i=1}^{d} \alpha_n \cdot f_n^i$$

（2）卷积层（Convolution layer）

Text-CNN 网络中通过卷积层来捕捉特征之间的关联性，论文中将特征因果图中的相邻特征放得很近，以便在卷积过程中可以捕捉到它们之间的关系。根据节点维度，作者将卷积核的大小设置为 3，4，5。

（3）全连接层（fully-connected layer）（这一部分查阅了相关资料）

在卷积层和池化层后，通常连接着 1 个或 1 个以上的全连接层，全连接是指每一层的神经元与其前一层的所有神经元进行全连接，全连接层每个神经元的激励函数一般采用 ReLU 函数。全连接层中最后一层的输出值传递给 softmax 层，采用 softmax 逻辑回归将实数范围内的分类结果转化为 0-1 之间的概率。

4.4 研究结论

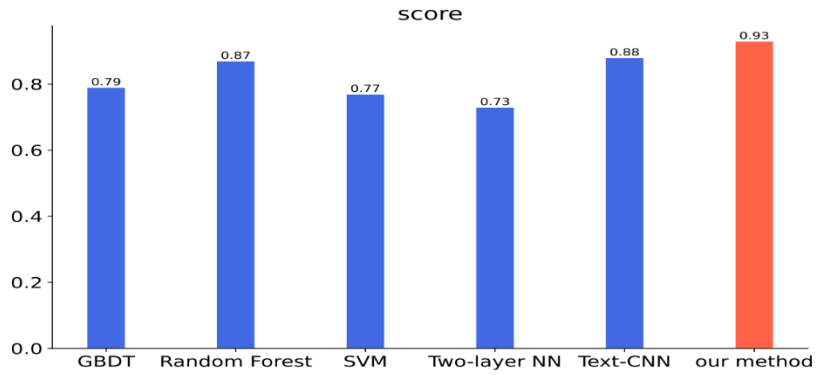

图 10 不同机器学习模型的测试得分

在介绍了 Text-CNN 模型后，作者还评估了多种机器学习模型的训练效果，如 GBDT,支撑向量机，随机森林。从上图中可以看出随机森林和 Text-CNN 两种方案明显的得分高于其他的模型。此外由于上述两种模型的测试结果相关性较低，作者将两种方案结合分别用于不同的根因分类，从而得出了最终方案：对根本原因 2 和 3 使用随机森林模型，对根本原因 1 使用 Text-CNN 分类，整体解决方案的得分为 0.93。

五、启发与思考

通过阅读这篇网络故障根因定位方面的论文，首先我学到了一些关于深度学习、文本序列分析方面的知识，初步了解了 Text-CNN 模型的构成和原理，扩充了自己的知识面。在论文阅读中我也体会到了作者的分析思路，在研究初期可以通过分析数据获得一些可用的结论，从而简化问题。此外针对问题难点设计方案，作者通过单因/多因判断，解决了多个根本原因同时出现的问题；通过 3 个二元分类解决了部分原因组合缺乏训练数据的难点，实现了化繁为简。我认为这篇论文的核心部分是后半部分数据处理的流程，在这一部分我也学到了一些可用的处理思路，例如如何将数据长度归一（补齐、截断）。此外，在阅读的过程中，我掌握了一定的阅读技巧，积累了阅读英文论文的宝贵经验。