一、 文章标题：

ACCURATE INFERENCE OF UNSEEN COMBINATIONS OF MULTIPLE ROOTCAUSES WITH CLASSIFIER ENSEMBLE

二、 想法：

我选择的论文是 ICASSP-SPGC-2022 AIOps 挑战的参赛论文，该文章对于无线网络中网络故障的根本原因分析问题提出了解决方案。通过分析特征分布特点，设计空间特征，对数据进行时间分析以确定原因个数，使用多个单一根本原因分类器的集成，并将 CNN 引入多元时间序列分类中，最终获得了较高的分类精度。通过这篇文章我对如何确定无线网络中网络故障的根本原因有了一定的了解，但更重要的是我学习了如何根据事件特征或错误情况来确定原因的方式。首先我们应当对错误进行定性定量的分析，发现并根据错误特征与错误原因之间的关联性，规定并提取有代表性的错误特征，选择合理的分类方法，以便获得良好的结果。

三、 阅读笔记：

准确的诊断和定位故障是确保移动网络可靠运行的基础，目前的研究方法主要有以下两种，分别是使用因果发现算法来推断时间序列之间的依赖关系和使用概率图形模型来表示变量之间的条件相关性，但这两种方法在无线通信中并未得到广泛应用。

无线网络故障的根本原因分析困难主要是因为无线网络故障具有特征类型多样；时间片长度不一，部分时间片长度过短；可能有多个原因同时存在，导致数据分布受到影响，从而影响对故障产生原因的判断；多种故障原因组合多样，实际应用中可能出现未知的组合，导致培训数据集不够全面。

在本文中，作者通过首先分析与根本原因高度相关的特征，其次通过对数据分布和故障原因的分析，作者发现当多个根本原因同时发生时，数据的分布会受到影响，基于此，提出了一种启发式方法来推断样本是包含单个根本原因还是多个根本原因。在学习各种机器学习模型的基础上，对多个单一根本原因分类器进行了集成。

针对上文所提出的难点，作者分别进行了解决。针对特征类型多样的问题，作者设计了特定的特征工程方法来表示所提供的空间特征；对于时间片长度不一和可能有多个原因同时存在的问题，作者通过结合了三个单一的根本原因二元分类器的集成方法成功实现了处理了不同长度的时间片和多个根本原因的未知组合；对于多种故障原因组合多样的问题，作者设计了高效的时间序列分析模块用以对存在多个根本原因案例进行分析，成功地预测多个根本原因的同时发生，最后，作者在根本原因推理任务中引入了从时间序列数据中找到根本原因的有效方法——TextCNN，进一步对算法进行了完善。

数据分析

2.1 相关性分析

作者绘制了随机选择样本的所有特征的相关矩阵，其中颜色越深，说明两个特征相关性越高，对比已知的因果关系，我们发现两个特征分布相关度越高，越可能是由相同的原因导致的，与观察结果相符，由此我们可以得到结论，特征相关度越高越可能具有共同的根本原因。

2.2 特征分布的方差

在这一部分，观察了不同根本原因下的特征分布。通过观察，发现对于某一根本原因而言，如果对于某一特征，该原因的分布与其他根本原因不同，则说明该特征是该根本原因的关键特征。同时也发现如果有多个根本原因同时发生，其分布特征与单个原因发生时差距较

大，如果将多个原因同时发生的特征分布作为其中一个原因的特征分布可能会导致分类器降级。

2.3 时间片分析

数据集由文件组成。每个文件都是一个时间片。时间片长短不一，但大多数文件长度都很短。我们通过在长文件中惠子特征的时间序列，认为长度近似的情况下，波动明显的片段可能具有更多的根本原因。因此，测量特征的波动，并在模型中考虑它们是必要的。

3.方法

3.1 框架

解决方案架构如图所示，首先分析样本的时间序列特征，以推断样本中是否存在多个根本原因的可能性。根据推理结果，遵循不同的分类策略进行分类。

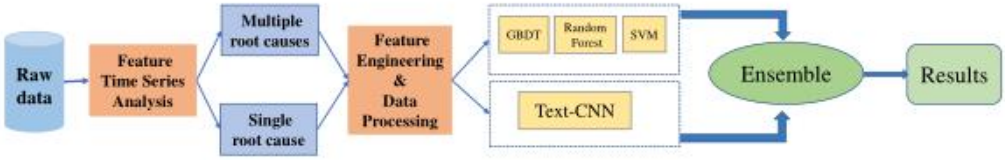

Fig. 4. Framework of our solution

算法一：分类策略

如果推断出一个样本中出现了多个根本原因，则对每一行数据进行根本原因分类，然后将所有行的分类结果进行集成，以获得最终的根本原因。反之，当推断结果为一个样本中只有一个根本原因，则对整个文件进行根本原因分类，以直接获得根本原因。

3.2 推断多个根本原因的发生

基于 3.2 中通过对片段波动性与根本原因个数相关联的设想，提出了一种方法来推断样本中多个根本原因的同时发生。首先根据现有因果关系图，选择与根本原因高度相关的特征。然后，在一个样本文件中评估它们的差异。如果差异性很大，则推断该文件存在多个根本原因。

3.3 特征工程与数据处理

3.3.1 空间特征设计

特征值的空间分布如图 5（a）所示，特征 20-n 表示八个接受方向的 ID，n 表示接受方向。取值为 0 到 31 之间的非负整数，排列为 4×8 矩阵。通过以下两个特征来使用空间特征：

（1）特征边缘：基于根本原因 2 指的是边缘区域的微弱信号，通过设计特征边缘来代表这一特征。首先将特征 20 矩阵的左两列和右两列定义为边列。然后，计算每个时间戳击中边缘列的特征 20 的八个方向值的数量，以形成特征边缘。

（2）特征距离：根本原因 3 指节点间的强干扰。考虑到接收方向之间的距离反映了信号之间的干扰程度，设计了特征距离。首先将矩阵放置在直角坐标系中，并将方向 ID 转换为坐标。然后，计算特征 8 个方向特征 20 的每两个节点在每个时间戳处的欧氏距离之和，以形成特征距离。

3.3.2. 特征长度设计

样本文件的数据长度差别很大。长度反映了不同根本原因的持续时间，将文件的长度计算为特征长度。

3.3.3. 其他特征

（1）Z-得分归一化：由于训练样本和测试样本之间的数据分布存在很大差异，因此需要执行 Z-得分归一化。

（2）数据扩充：由于给定的标签极不平衡，因此使用边界 SMOTE 算法进行数据扩充。

3.4 分类器

3.4.1. 基于 TextCNN 的分类器

文章使用修改后的 TextCNN 来捕捉动态特征之间可能的因果关系。给定一个由几十个特征序列组成的时间片后，从特征本身和特征之间的因果关系来推断根本原因。因此，我们使用 TextCNN 获得动态特征来代表特征之间的因果关系。然后，在给定因果图的情况下，将特征路径，即图中从起始特征节点到结束特征节点的路径作为判决，将特征的时间序列作为特征向量。

（1）填充剪裁：由于 TextCNN 中单词向量的长度是固定的，所以需要对文件特征值进行填充和剪裁。由于过半的文件长度在 30 左右，因此，对于长度超过 30 的文件，剪裁数据；对于小于 30 的，使用平均值来填充特征。然后通过相邻两个文件的线性插值填充所有列。最后，每个文件形成一个矩阵[number of Feature×30]作为输入。

（2）注意层：为了对根本原因进行分类，不同的特征可能具有不同的重要性。因此引入了一个注意层来动态计算特征权重。注意层的计算结果为：

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

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

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

其中，W_n^T是一个可训练参数，f表示特征向量，α_n是每个特征的注意权重，h_n是注意层的输出。

（2）卷积层：在注意层之后，将相邻特征放在通过 CNN 捕捉特征之间的因果关系，以便在卷积过程中捕捉它们之间的关系。卷积核的大小是 TextCNN 捕捉特征之间相关性的一个重要参数。由于因果图中节点的度数大多为 3、4、5。因此，将卷积核的大小设置为 3、4、5。在汇聚层之后，我们使用完全连接层和 softmax 层进行分类，并输出每个类别的概率。

在汇聚层之后，使用完全连接层和 softmax 层进行分类，并输出每个类别的概率。

3.4.2. 机器学习分类器集成

除了 TextCNN，该文系统地评估了其他经典机器学习模型，包，并在测试数据集上绘制了不同算法的评估结果。基于竞赛平台的分数，同时考虑到 Random Forest 和 TextCNN 之间的相关性很低，使用这两个模型在数据中识别不同的根本原因。参考 One vs Rest 策略，使用三种不同的单根本原因二进制分类器。最终使用 Random Forest 作为根本原因 2 和 3，使用 TextCNN 作为根本原因 1。