**清华大学-中国移动联合研究院**

**科研课题立项申请书**

**课题名称：网络智能感知和诊断技术**

**申报单位：九天智能研究中心**

**2021年4月**

**目录**

[课题信息表 3](#_Toc63852209)

[课题立项申请书 4](#_Toc63852210)

[一、需求分析 4](#_Toc63852211)

[1、国家、行业或中国移动在技术发展、业务应用发展方面存在的需求 4](#_Toc63852212)

[2、存在的科学技术问题 4](#_Toc63852213)

[3、现有相关研究基础与局限 5](#_Toc63852214)

[二、研究内容 7](#_Toc63852215)

[1、研究目标 7](#_Toc63852216)

[2、主要研究内容 7](#_Toc63852217)

[3、创新性和可行性分析 8](#_Toc63852218)

[三、 实施方案 9](#_Toc63852219)

[**1、技术路线、研究方法** 9](#_Toc63852220)

[2、课题研究周期及重要节点 12](#_Toc63852221)

[3、风险分析及控制方案 13](#_Toc63852222)

[四、预期成果 14](#_Toc63852223)

[1、课题成果描述 14](#_Toc63852224)

[2、重要节点考核指标 14](#_Toc63852225)

[2021年度课题实施方案 16](#_Toc63852226)

[一、科学问题及主要研究内容 16](#_Toc63852227)

[二、 课题实施计划 17](#_Toc63852228)

[三、 职责分工 18](#_Toc63852229)

[四、课题成果呈现形式及验收方法（2021年） 18](#_Toc63852230)

[附件1：2021年度人员投入情况 19](#_Toc63852231)

[附件2：年度课题经费预算（详见《课题预算申报表》） 25](#_Toc63852232)

# 课题信息表

|  |  |
| --- | --- |
| 课题名称 | 网络智能感知和诊断技术 |
| 研究方向 | 平台关键AI能力 |
| 课题  负责人 | 负责人1（清华大学）姓名：裴丹 单位及职务/职称：清华大学计算机科学与技术系/长聘副教授  负责人2（中国移动）姓名：朱琳 单位及职务/职称：中国移动研究院人工智能与智慧运营中心技术经理/高级工程师  负责人3（中国移动）姓名：刘子靖 单位及职务/职称：中国移动研究院业务研究所项目经理/工程师 |
| 课题周期 | **2021年5月 – 2023年12月** |
| 预期成果 | 2021-2023年，预期完成 “适应性NFV日志压缩方案”、“无线网多指标异常检测算法”、 “家宽场景的根因定位”、“家宽场景的异常预测”、“运维知识图谱”等五个研究内容的方案，分别对应一套原型系统产出，在中国移动的NFV、无线网和家宽等场景实际部署应用，支撑现网运维，并在每个研究内容产出一篇高质量学术论文。相关模型承载到网络智能化平台，提供开放调用服务；家宽异常检测等模型承载于家宽业务和网络质量监测系统（软探针系统）在现网部署。预期共计产出5个原型系统，5篇学术论文。 |
| 预期技术  标准制定 | □国际标准 □国家标准 □行业标准 □企业标准 |
| 课题周期内人员投入 | 清华大学：35人参与，投入630人月 |
| 中国移动：12人参与，投入90人月 |
| 课题经费 | 课题总经费： 1200 万元  2021年度课题经费： **218.52 万元** |

# 课题立项申请书

## 一、需求分析

### 1、国家、行业或中国移动在技术发展、业务应用发展方面存在的需求

中国移动服务产品丰富，如4G网络、家庭宽带网络，以及正在大规模部署推广的5G网络，网络场景多样、结构复杂、数据多样且服务的用户量庞大。随着网络在人们生活中的重要性日益增长，用户对服务质量的要求也越来越高，这些都给网络产品的运维增加了难度和压力。传统的运维模式，以较多人力参与为主，运维效率以及准确性上都已逐渐无法满足需求，在当前降本增效提质的压力下，急需引入AI方法来为运维赋能，网络智能感知和诊断是业界普遍的重要需求。

近年来，以机器学习和深度学习为代表的网络感知和诊断在网络异常检测、关联分析、根因定位、日志解析等关键问题上取得了突破性的进展，将人工智能技术引入网络运维，可以为运维场景提供高准确性和及时性的服务，达成智能运维的目标，进一步解放人力，提高运维效率，并提升用户感知。

中国移动的运维场景多样，逐一场景攻关工作量过大，需要聚焦共性需求，以标杆应用为依托，打通从数据收集处理，到网络感知和诊断的端到端闭环流程，沉淀核心能力，进而支撑众多应用，实现规模价值。梳理的标杆应用包括：NFV日志的模式发现和压缩、无线网场景的多指标异常检测、家宽场景的故障根因定位、家宽场景的异常预测、运维知识图谱。

### 2、存在的科学技术问题

由于通信网络的特殊性，造成当前的机器学习或者深度学习不能很好的适用于通信网络，具体包括：

1）通信网络的网元或者设备种类多样、数量众多，各个网元或者设备间的数据相差很大，以NFV日志为例，不同厂家不同网元不同设备的日志均存在较大差别。以自然语言处理的摘要提取等实现的日志解析算法，虽然可以适配不同的设备，但解析的结果不准确；以模板提取等实现的日志解析算法，虽然对具体设备效果很好，但是很难适配各种不同的设备。如何设计通用的日志模式发现和压缩模型，进而支撑后续的数据分析等，在实际应用中尤为重要。

2）通信场景复杂多变，运维数据海量且相互间关系复杂，当前基于机器的异常检测算法虽然可以解决海量数据问题，但对多指标的复杂时序性和关系性表达能力不强；基于深度学习的异常检测算法虽然可以对多指标时序数据很好的表达，但是由于需要过多的训练数据，很难应用到海量场景中。如何设计一种适用于海量数据的多指标时序数据异常检测算法，是一个亟待解决的挑战性问题。

3）通信网络场景的特殊性，已有的经验无法很好的适用于通信场景，例如在家宽领域的根因挖掘中，由于家宽网络具有业务长流程、端到端环节多、家庭内网环境复杂，质差异常监测难、根因定位分析难的特点。根据传统经验可能最先发生的异常指标是根因、与异常指标相关性最大的指标是根因，但是由于家宽场景的特殊性，承载业务多样化、网络条件复杂，这些经验均不符合要求，需要针对家宽领域的特殊性，从网元多样化业务质差检测特征选择（关键指标）、时间序列突变检测机制、时序随机游走路径设定等方面进行模型设计和算法调优，形成理论突破和算法创新。

4）中国移动家宽业务用户数量庞大，尽管业务发生故障概率较低，但总用户投诉数量依然较大。目前缺乏较为准确的故障预测方法，以期能够在用户投诉前，主动预测哪些用户发生故障的概率较大，并主动采取措施规避故障，这将极大减少用户投诉数量，提高用户满意度。

5）中国移动家宽业务监控数据类型多样，数量庞大，且规律各异。传统的运维方法针对各个场景逐一独立解决，这一过程往往耗费大量的人力物力资源，效率低下，数据及算法的共享性非常低。近年来，知识图谱研究发展迅速，但运用在智能运维场景中的知识图谱研究还不够成熟，缺乏大规模高效可用的运维知识图谱的构建理论与方案。

### 3、现有相关研究基础与局限

**任务1：根据现网实际需求，研究准确、高效的NFV运行日志模式发现算法及面向实时检索的日志压缩算法。**

**研究基础：**日志是通信网络设备运行维护最重要的信息源，保留了运维过程的原始信息，可以基于日志数据完成设备的异常发现、检测和定位等操作，但是随着网络设备和服务规模的不断增长，日志数据量越来越大，直接基于原始非结构化的日志数据分析不现实，需要有一种合适的方法来对大量的非结构化文本日志数据做压缩。

日志是由通信软硬件系统中的“printf”函数生成的非结构化文本。一条日志通常具有固定的结构，这一结构包含了时间戳、软硬件系统ID、消息类型和详细消息5个域。其中，时间戳域表示日志生成的具体时间；软硬件系统ID域表示生成日志的软硬件系统的标识；消息类型域描述了日志的概要特征；详细消息域描述了日志的具体事件。一般的，消息类型域和详细消息域的语法随着软硬件系统的类型、生产厂商、型号的变化而变化，没有统一的格式，但通常由固定部分和参数部分组成。固定部分是开发人员预定义的、用于表示某类事件的信息，参数部分在软硬件系统运行过程中依据具体的时序、交互设备等信息动态生成。

**已有研究的局限：**在自然语言处理领域有一些关于文档摘要、语句压缩和聚类的方法。然而，有文献证明了自然语言处理的方法不能准确的解析日志。为了实现自动日志解析的目标，学术界提出了许多快速准确的数据驱动模板提取方法，例如Spell，LogSig和IPLoM。然而，现有的日志解析方法并不具有适应性，这种适应性表现在两个方面，即服务内的适应性和跨服务的适应性。没有适应性的日志解析方法极大限制了包括日志压缩在内的诸多应用。

**任务2：无线网多指标时间序列异常检测的通用性方案研究。**

**研究基础：**真实世界运行的系统，例如通信无线网络各网元产生的时间序列数据规模正在急剧增长，从这些为监控系统运行情况而采集的数据中推断出特定时刻系统的异常是对系统进行及时的维护和修复所必须的。例如，如果能从无线网网元输出数据中及时发现网络设备故障，就有可能避免故障引发的损失。

其中，一个指标的监测数据形成了一个单指标时间序列，具有多个指标的网元监测数据就形成了一个多指标时间序列。多指标时间序列异常检测十分重要，但也面临许多困难。比如：

1) 多指标时间序列异常检测比单指标时间序列异常检测要更复杂，有些情况下，虽然出现单个指标异常，但可能系统表现是正常的。

2) 多指标时间序列规模庞大，数量巨多。

3) 不同时段、地理位置的网络场景的异常检测模型需要适应。如何快速迁移并重训练原有模型是有挑战的。

现有的解决方案可以分为以下几类：

多元时间序列的异常检测是一个快速发展的领域。在深度学习方法得到广泛应用之前，传统的异常检测方法可以分为三类: 基于某种距离度量计算样本间距离，例如 kNN；基于无监督聚类；估计正常样本分布的概率密度，例如 One-Class SVM。

由于聚类通常也是通过计算样本间的某种距离度量来完成的，第一类思路和第二类思路的区别只是在对未知数据进行推断时是否需要保留训练样本，使用聚类方法的话，保留的是训练得到的模型参数。DCN 延续了这两类思路，通过结合 AutoEncoder 和 k-means 来同时获得高质量的降维和低维聚类结果，换言之，它通过训练神经网络得到用于聚类的距离度量。而 DAGMM 则延续了第三类思路，通过结合 AutoEncoder 和 GMM 模型，来同时获得高质量的降维和低维分布概率密度。

对时间序列中的时序关系进行建模也是一个经典的问题，传统方法包括自回归滑动平均模型 (ARIMA) 等。深度学习领域中则通常使用 LSTM 神经网络层来对捕捉时间序列中的时序依赖关系。

**已有研究的局限：**基于深度学习方法虽然可以很好的捕捉时间序列中的时序依赖关系和多时间序列的复杂关系，都需要比较长时间的训练数据，才能够很好的训练出一个优秀的模型。但这与现实情况是矛盾的，实际情况中，为了修复错误、完善功能、提高性能等，网络设备的部署或更新是十分频繁的。这代表着多指标时间序列的数据分布也会发生剧烈变化，新产生的数据集和训练集具有不同分布，如果继续使用之前训练好的模型，就会带来大量的误报。且由于对较长时间的训练数据要求，深度学习模型很难支撑海量数据的应用。

**任务3：针对家宽网络服务的智能运维场景**

**家宽子任务1：家宽网络场景的故障根因快速定位。**

随着网络转型的升级，家宽业务长流程，故障监测、定位难的特点，家宽业务运维难度不断提升，传统的以人工操作为主、机器自动操作为辅的运维方式，难以高效高质地支撑家宽业务运维。随着AI技术飞速发展，AI逐渐为各行业带来了创新发展新机遇，对于数字化、信息化程度较高的通信行业，AI将会为通信行业的数字化服务创新转型带来新的机遇，同时，业务探针产品生成的海量数据，共同为解决现网家宽业务运营共性痛点问题提供了技术和数据基础，因此需要研究如何在家宽业务运营及运维过程中引入AI技术，并进行建模和产品化工作，以支撑家宽应用需求，提升网络运营运维效率、降低网络运维成本。例如：用户常访问网站的故障监测及告警诊断功能，针对节假日，用户上网的高峰时段等，可以快速发现用户上网问题，并准确诊断到问题网元，从而快速进行干预处理，减少用户投诉。

现有相关研究基础与局限：

基于目前已有的研究工作发现，大规模系统的进行故障根因定位包含两个主要部分：学习组件间的关系去构建故障传播图和通过随机游走推断根因。但是这种想法在家宽业务中面临三个主要挑战：

1）基于独立且均匀分布（表示为iid）数据的故障传播图算法无法学习家宽场景时间序列之间的顺序关系。当前生成故障传播图通用算法PC是为iid数据设计的。因此，他们将每个时间点的数据视为相互独立的样本，从而忽略了时间序列中基于时间先后的时序性，导致故障传播图缺少关键的边，而无法用于根因定位。

2）与异常指标的相关性无法找到根本原因。基于随机游走的假设，与指标异常最相关的指标更有可能是根本原因。但是，在我们的方案中，监视指标是异类的。来自不同类别的监控指标的相关性低于来自相同类别的监控指标。因此与指标同类的指标容易被误判为根因。

3）监控指标的顺序关系和优先级不能用于随机游走。由于故障传播的原因，首先出现异常的指标可能是根本原因。由于数据的聚集，我们可以观察到某些指标同时出现异常。因此，我们不能仅基于顺序信息来定位根本原因。此外，不同的监控指标反映了家宽业务系统中的不同部分，它们在故障传播过程中具有优先权。在先前的随机游走中，不能使用监视指标的顺序关系和优先级。

4) 故障传播图建立的过程中，探索图知识图谱在解决该问题方面的优势。通过分析复杂监测指标之间的拓扑关系，初步建立指标指标之间的知识图谱模型。

**家宽子任务2：建立家宽故障的预测模型，能先于用户投诉发现问题。**

中国移动家宽业务用户数量庞大，尽管业务发生故障概率较低，但总用户投诉数量依然较大。如果能够在用户投诉前，主动预测哪些用户发生故障的概率较大，并主动采取措施规避故障，将极大减少用户投诉数量，提高用户满意度。为实现家宽业务的故障预测，需要重点解决监控数据规模巨大、多样化家宽业务监控数据特征复杂、故障样本差异较大的挑战，。此外，监控数据样本存在严重失衡，即故障样本数量远低于正常样本数量，需要对现有预测模型基于家宽故障小样本数据进行优化设计，提升模型的准确率。

**家宽子任务3：中移动场景下故障分析相关的知识图谱。**

中国移动家宽业务监控数据类型多样，数量庞大，且规律各异。针对运维的每一个场景，如异常检测、故障定位、根因分析、故障预测、故障规避、容量调度等，往往都需要经过数据清洗、数据填充、特征抽取、规律学习的过程。这一过程往往耗费大量的人力物力资源，效率低下。近年来，知识图谱研究发展迅速，但运用在智能运维场景中的知识图谱研究还不够成熟，缺乏大规模高效可用的运维知识图谱的构建理论与方案。

## 二、研究内容

### 1、研究目标

本课题面向中国移动通信网络中存在的若干运维相关科学问题，通过对智能运维领域的前沿技术理论的深入研究，并聚焦网络共性需求，以标杆应用为依托，打通从数据收集、处理、到网络感知和诊断的端到端闭环流程，沉淀有针对性的、高效普适的智能感知与诊断技术方案，进而支撑众多应用，实现规模价值。

具体为以下子目标：通过对通用日志模板提取技术与日志压缩技术的探索，实现适应性NFV日志压缩方案；通过探索多时序数据的聚类及异常检测技术，实现无线网多指标异常检测方案；通过对故障传播图技术的研究，实现家宽网络服务场景中的故障根因定位方案；通过对基于有监督学习算法的预测算法研究，实现家宽场景的异常预测；通过对知识图谱技术的研究，实现中移动场景下故障分析相关的知识图谱方案。

### 2、主要研究内容

**任务1：设计并实现准确、高效的NFV运行日志模式发现算法及面向实时检索的日志压缩算法。**

本任务需要克服现有的日志解析方法适应性方面的两个缺陷：服务内的适应性和跨服务的适应性。研究内容包括：首先探索通用日志模板的提取算法，实现对日志的模板的训练提取，然后利用训练好的模型匹配在线日志、并能够增量式地更新模板集合。最终实现适应性NFV日志压缩系统，在中国移动省公司真实NFV场景部署应用。

**任务2：无线网多时间序列聚类及异常检测的通用性算法研究。**

在无线网络监测场景中，大规模网元不间断地生成各类监控数据，这些监控数据构成了海量的多指标时间序列。针对海量多指标时间序列异常检测的通用性算法研究，在中国移动的4G，尤其是规模商用的5G等领域的智能监控，具有广阔应用场景。

本课题的研究内容包括：首先拟研究海量的多指标数据基于形状的聚类方法，使得同一类内的多指标数据形状相似；然后对于同一类内的数据提取中心数据，学习和训练一个共有的无监督多指标异常检测模型，适配于类内的多指标时序数据；最后优化模型的时效性，保证模型可以嵌入到中国移动的多指标时序数据异常检测系统，服务中国移动省公司的无线网络监测场景。

**任务3：针对家宽网络服务的智能运维场景**

**家宽子任务1：针对家宽网络服务场景，实现故障根因快速定位方案。**

大规模的家宽业务系统通常有多种不同的业务和系统指标实时监控系统运行情况。本课题研究内容包括：基于对故障时间点前的一段数据进行分析，首先建立监控指标之间的故障传播图模型，其次，对诸多监控指标的进行异常检测，获取包含异常程度、异常时刻等信息，然后采取随机游走算法对建立的故障传播图进行试验，得出潜在的根因指标，最后再结合异常信息与潜在根因信息，得出异常根因的推荐排序。

**家宽子任务2：建立家宽故障的预测模型，实现先于用户发现异常问题。**

本课题研究内容包括：基于对最新网络故障预测技术的全面跟踪梳理，拟采用基于监督学习算法的预测算法，设计高准确性、低漏警率、高通用性的针对中移动特定场景下的网络故障预测方案，并产出相关技术文档、原型系统等成果。

**家宽子任务3：实现中移动场景下故障分析相关的知识图谱。**

本课题拟探索智能运维知识图谱技术，将家宽业务监控数据进行有机整合，建立知识图谱。本课题拟研究内容包括：首先研究基于数值指标数据的特征提取方法，以及日志文本类数据的关键信息提取技术，其次研究前沿的运维知识图谱建立方法，然后统筹多类型数据，实现新型实体发现、新型关系发现、新型属性发现等技术，从而构建高实用性的智能运维知识图谱。

### 3、创新性和可行性分析

**1）创新性：**

**任务1：对于NFV运行日志模式发现及日志压缩。**本课题拟实现基于异构日志的通用模板提取方法，并基于少量数据训练准确的日志解析，拟解决原来的日志解析方法由于新部署服务中日志量不足而不能够准确学习新模板的问题。

**任务2：无线网多指标时间序列的聚类分析及异常检测研究。**本课题拟使用随机变量连接以挖掘模型的潜在表示，创新性地从分布更复杂、历史周期更长的数据中学习多指标的正常模式。本课题拟通过结合基于聚类+异常检测的算法，在保证较高检测准确度的前提下，大大降低了模型的训练开销。算法将应用于中国移动无线网络的运维场景，解决多指标时间序列的聚类和异常检测问题。

**任务3：家宽网络服务的智能运维场景**

**家宽子任务1，面向家宽场景的根因定位。**本课题拟设计PCTS算法构建时间序列的故障传播关系，在随机游走算法中采用偏相关算法来追溯根因；拟通过结合故障传播图学习算法和随机游走算法，无需过多人工设计，通过分析数据间的关系，自动的定位系统根因，实现较高的定位准确率。

**家宽子任务2，面向特定场景的故障预测方法研究**。基于大量监测数据，采用基于有监督机器学习的方法，挖掘故障规律，拟实现在用户投诉前，主动预测哪些用户发生故障及引发投诉的概率较大。故障预测问题难点在于准确性普遍偏低，本课题拟实现高准确性的预测方案，代替基于有监督机器学习的方法。该研究具有较强的创新性和实际价值。

**家宽子任务3，构建运维知识图谱。**以往知识图谱通常是针对同一类数据信息进行构建，本课题拟结合不同类型数据信息，如数值指标、日志文本、故障事件等信息，将家宽业务监控数据进行有机整合，构建高实用性的智能运维知识图谱。

以上研究内容将面向中国移动大规模网络中的运维场景进行强针对性的算法研发。算法方案除了具有一定普适性外，将对中国移动特定运维场景进行调整与适配，以期能在中国移动大规模复杂网络中具有良好的效果。

**2）可行性：**

以下将从已有研究基础方面、研究队伍和实验条件方面，对本课题的可行性进行分析。

**针对任务1**，本项目预计使用的场景拥有大量的非结构化NFV日志，申请人团队已经通过前期对传统离线日志解析算法的调研的测试，发现传统的算法在离线场景下可以取得良好的效果，所以在设计本项目时，使用传统算法提取离线模板，然后为在线的日志解析提供机器学习的训练数据集。这样不仅解决了数据无标注的问题，并且能够因为使用准确的训练数据集而提高在线模型的准确率。

**针对任务2**，在已有研究基础方面，申请人先后在计算机领域的国际期刊和会议（包括IWQOS、KDD）上发表了关于时间序列聚类和多时间序列异常检测相关的文章：“Zhihan Li, Youjian Zhao, Rong Liu, Dan Pei. Robust and Rapid Clustering of kpis for Large-Scale Anomaly Detection. IEEE/ACM IWQoS 2018, Ban , Alberta, Canada, June 4-6, 2018 (CCF 推荐 B 类会议)”，“Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, Dan Pei. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. ACM SIGKDD 2019, Anchorage, Alaska, USA, August 4-8 2019 (CCF 推荐 A 类会议)”。

**针对家宽场景的子任务1**，在已有研究基础方面，申请人团队已在计算机领域的国际期刊和会议IWQOS上发表了关于大规模计算机系统根因定位相关的文章：“Yuan Meng , Shenglin Zhang, Yongqian Sun, Ruru Zhang , Zhilong Hu, Yiyin Zhang, Chenyang Jia, Zhaogang Wang, Dan Pei. Localizing Failure Root Causes in a Microservice through Causality Inference. IWQOS2020,(CCF 推荐 B 类会议)”.

**针对家宽场景的子任务2和3**，本团队具备互联网服务提供商智能运维数据分析的经验，熟悉家庭宽带业务的数据类型和规律，在互联网内容提供商、商业银行等企业的智能运维有丰富的实践经验。且具有一定的理论基础，如已发表相关文章：“Shenglin Zhang, Ying Liu, Weibin Meng, Zhiling Luo, Jiahao Bu, Sen Yang, Peixian Liang, Dan Pei\*, Jun (Jim) Xu, Yuzhi Zhang, Yu Chen, Hui Dong, Xianping Qu, Lei Song. “PreFix: Switch Failure Prediction in Datacenter Networks “. ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems 2018, Irvine, California, USA, June 2018 (CCF 推荐B类会议)”，“Shenglin Zhang, Weibin Meng, Jiahao Bu, Sen Yang, Ying Liu, Dan Pei, Jun (Jim) Xu, Yu Chen, Hui Dong, Xianping Qu, Lei Song. “Syslog Processing for Switch Failure Diagnosis and Prediction in Datacenter Networks”, IEEE/ACM International Symposium on Quality of Service (IWQOS), VILANOVA I LA GELTRÚ, SPAIN, June 2017 (CCF 推荐 B 类会议)”等。

## 实施方案

**1、技术路线、研究方法**

**任务1：设计并实现准确、高效的NFV运行日志模式发现算法及面向实时检索的日志压缩算法。**

本任务首先要克服现有的日志解析方法适应性方面的两个缺陷：服务内的适应性和跨服务的适应性。对于服务内部而言，需要实现模板自适应更新；对于跨服务的适应性，需要实现，对于能自动适应不同服务器NFV日志的通用性模板。

我们首先将日志转换成为“模板编号+变量”（例如“T1+ae3”），这样就节省了大量的存储空间。基于模板的日志压缩也需要对任何给定的日志提供相应的模板。换句话说，它也需要一种适应性的日志解析方法。

就NFV各网元内部适应性来说，我们使用任何现存的传统模板提取方法得到单词标签。然后，我们利用训练好的本项目模型匹配服务器A的在线日志。如果匹配失败，它将会学习新的模板并且增量式地更新模板集合。对于NFV跨网元的适应性来说，我们直接使用网元A的日志训练得到本项目的模型，然后去匹配网元B的在线日志。

**任务2：设计并实现无线网多时间序列聚类及异常检测的通用性方案。**

本课题首先对无线网大规模网元的多指标数据进行基于形状的聚类，同一类内的多指标数据形状相似；然后对于同一类内的数据提取中心数据，学习和训练一个共有的无监督多指标异常检测模型。因此，本课题主要分为两部分技术：多指标数据聚类和无监督多指标异常检测。

1. 指标聚类技术方案

指标聚类算法的设计，主要包括三个部分。

* 数据预处理与基线（baseline）获取。

对于不同机器的指标数据，需要首先经过预处理以统一量纲。此外，指标曲线的异常与噪声会干扰曲线的相似性判断。因此，本项目设计算法获取每条指标曲线消除噪声后的基线，作为聚类算法的输入。

* 相似性判别。

基于互相关(cross-correlation)定义曲线间的相似性距离，主要考虑不同曲线在形状上的相似性，以此作为聚类算法中的距离度量。

* 聚类与分派(assignment)算法。

考虑到指标曲线的类别不均衡性以及可能存在的相位偏差，采用基于密度的算法进行聚类。同时，改进该算法以自适应地决定算法参数，也可以通过人工调整得到不同精细粒度的聚类。对于得到的每个聚类簇，给出代表该类别的聚类中心，方便运维人员了解各个类别的特征。面对大规模指标曲线数量带来的巨大计算开销，随机采样一小部分进行聚类，对其余曲线利用聚类中心进行快速分派，最终给出每条指标曲线所属的类别。

在此基础上，对机器进行聚类，并确定每一类的类别中心机器。

1. 多指标异常检测技术方案

多指标异常检测的基本思想是首先使用GRU捕捉原始多指标数据间的复杂时间依赖关系，并结合一种常用的算法——VAE将输入变量映射到随机隐藏变量。其次，为了明确地建立随机隐藏空间中随机变量之间的时间依赖关系，提出一种随机变量连接技术：使用线性高斯状态空间模型（SSM）连接随机变量，并将随机隐藏变量与GRU隐藏变量进行拼接。最后，为了帮助变分网络中的随机变量捕获输入数据的复杂分布，采用planar NF学习随机隐藏空间中的非高斯后验分布。

多指标异常检测方法整体可以分为两个部分：离线训练部分和在线检测部分。数据预处理是离线训练和在线检测共享的模块。在数据预处理过程中,首先对数据集进行标准化处理，然后通过一定长度的滑动窗口将数据分割成多组序列片段（模型训练和检测的输入是一段时间序列）。经过预处理后，训练数据被发送到模型训练模块。

模型训练模块的目的是学习捕获多指标数据的正常模式并为每个数据点输出异常分数。阈值选择模块使用离线训练时的异常分数，根据POT（Peaks-Over-Threshold）方法自动选择异常阈值。离线训练模块可以定期执行（例如每周或每月进行一次）,从而适应最新的数据特征。

根据已经训练好的模型，预处理后的新数据点将输入到在线检测模块，得到其异常得分。如果该时刻的异常得分高于异常阈值，则将该时刻判断为异常，否则为正常。如果检测某时刻发生异常，通过评估该时刻每个指标的贡献（即重建概率）来解释这次异常。

**任务3：针对家宽网络服务的智能运维场景**

**家宽子任务1：针对家宽网络服务场景，涉及并实现快速定位故障根因的方法。**

本课题旨在设计一个自动识别异常根因的框架。结合家宽业务系统的特点和实际的需求。当在系统中检测到KPI发生异常时，根因定位框架根据经验性故障传播时间，从故障时间之前的一段时间到此刻为止的监视指标用作算法的输入。输入数据集将用于生成故障传播图。同时，对数据集中的指标进行异常检测。故障传播图学习模块和异常检测模块可以并行处理。然后，基于故障传播图和指标的异常信息，随机游动将给出故障的前N个潜在根本原因。

1. 故障传播图学习

我们拟提出升级版PC算法，基于监控指标学习故障传播图，其中每个节点代表指标的一个时间点。

1. 异常检测

在根因定位系统中，我们假设根因指标应该在故障发生之前的某个时间变得异常。拟采用极值理论来检测指标的异常。对于时间序列中的每个时间点，极值理论将基于过去数据的极值分布生成一个阈值。高于高阈值或低于低阈值的时间点将被视为异常。

1. 时间序列的随机游走算法

我们拟设计一种时间序列的随机游走算法，对潜在的根因指标进行排序，并给出前N个根因。

**家宽子任务2：建立家宽故障的预测模型，能先于用户发现问题。**

本研究期望能够在用户投诉前，主动预测哪些用户发生故障的概率较大，并主动采取措施规避故障，这可极大减少用户投诉数量，提高用户满意度。

本课题原理基于一项观察，故障的出现通常是由量变积累到质变的过程，故障暴发之前已经出现一些先兆，这些先兆可以通过对于数据的分析挖掘进行建模。首先对最新网络故障预测技术的全面跟踪梳理，拟采用基于监督学习算法的预测算法，再通过对大量故障时刻上下文数据进行学习训练，挖掘故障发生前的数据变动规律，建立故障预测模型；然后再基于具体场景中实例优化调整预测模型，使其达到在该场景中具有较高的预测准确度。

**家宽子任务3：中移动场景下故障分析相关的知识图谱。**

传统的智能运维研究思路，往往针对运维的每一个场景，如异常检测、故障定位、根因分析、故障预测、故障规避、容量调度等，往往都需要经过数据清洗、数据填充、特征抽取、规律学习的过程。这一过程往往耗费大量的人力物力资源，效率低下。

本课题拟基于前言的智能运维知识图谱技术，近年来，得益于人工智能的快速演进，知识图谱研究得以迅猛发展。首先研究基于基于数值指标数据的特征提取方法，以及日志文本类数据的关键信息提取技术，然后再结合多种智能运维异常发现、异常定位技术，发掘异常及故障事件，最后统筹多类型数据，实现新型实体发现、新型关系发现、新型属性发现等技术，从而构建基于多源数据的，信息全面、高实用性的智能运维知识图谱

### 2、课题研究周期及重要节点

课题研究周期为2021年至2023年，重要节点包括：

1. 2021年6月：

* 实现家宽场景根因定位算法初步设计。
* 提出NFV日志解析初步设计方案。
* 无线网指标数据聚类方法初步设计

1. 2021年8月：

* 完善家宽场景根因定位算法设计。

1. 2021年10月：

* 家宽场景根因定位算法的原型系统初步完成。
* 实现并优化适应性NFV日志解析方案的初步方案。
* 无线网多指标数据聚类方案原型系统的实现。
* 无线网无监督异常检测算法完成初步设计。

1. 2021年12月：

* 家宽场景根因定位算法的原型系统初步完成。使用中国移动家宽场景的真实数据完成原型系统验证，预计准确性达到60%。完成算法代码方案的实现和效果评估、测试，优化系统的整体性能。开始筹备1篇论文投稿。
* 实现适应性NFV日志压缩方案原型系统初步完成。使用中国移动省公司真实数据完成验证，使得模板提取准确率达到60%。
* 基于无监督机器学习的多指标异常检测算法原型系统初步完成，并与多指标数据聚类方法进行有机整合，进一步优化系统的整体性能。基于中国移动省公司无线网数据初步完成验证，综合准确性达到约80%。完成1篇相关论文投稿。

1. 2022年~2023年：

* 预期将进一步完善以上三个研究任务的原型系统及论文，并实现在中移动实际的家宽、NFV和无线环境中的部署应用。
* 家宽场景根因定位算法的原型系统测试原型系统进一步完善，Top5的定位准确性达到85%以上，支撑中国移动家宽场景的部署应用。发表1篇相关论文。
* 适应性NFV日志压缩方案原型系统进一步完善，使得模板提取准确率达到85%，原始日志与压缩后的日志的压缩率达到30%，支撑中国移动不少于1个省公司NFV场景的部署应用。发表2篇相关论文。
* 基于无监督机器学习的多指标异常检测算法原型系统进一步完善，综合准确性达到约85%，支撑中国移动不少于1个省公司无线网场景的部署应用。发表1篇相关论文。

### 3、风险分析及控制方案

**项目可能风险包括：**

1）模型在线推理风险：现网各网元或设备间数据分布不同，可能导致离线训练的模型，部署时线上模型和离线模型效果不一致，需要针对线上部署情况优化模型。

2）算法实时性风险：算法实时性要求高，但离线训练追求效果提升，可能导致实时性不高。所以需要保证模型效果的同时，提升时效性。

3）数据规范性风险：现网数据规模大且样本多样，且可能存在数据不规范的情况，模型需要适配各种可能数据。

4）模型先进性的风险：现网场景多变，如何保持模型的自更新和自迭代。

5）算法部署风险：大规模部署可能对算力要求较高。

6）项目进度风险：项目进度可能不满足要求。

**控制方案：**

1）模型在线推理风险的控制：整理不同网络不同网元等各种数据，提升模型的通用性。

2）算法实时性风险的控制：调研模型压缩或优化的方法，提升模型实时性。

3）数据规范性风险的控制：调研模型优化方法，提升模型通用性，支撑各种类型的数据；同步优化数据情况，提升数据的不规范性。

4）模型先进性的风险的控制：调研模型自更新方案，保持模型先进性。

5）算法部署风险的控制：通过模型通用性、模型压缩、集中部署等方案，提升模型大规模部署能力。

6）项目进度风险的控制：项目按照月报和阶段里程碑跟进控制，降低项目风险。