北京大学
PEKING UNIVERSITY

硕士研究生学位论文

题目：AETA 原始数据分析平台的设计与实现

姓　　名：	马滨延
学　　号：	1701213535
院　　系：	深圳研究生院
专　　业：	微电子学与固体电子学
研究方向：	系统集成芯片(SOC)设计及设计方法学
导师姓名：	王新安 教授

二〇二〇　年 六 月

版权声明

摘要

地震是一种给人类生命和财产安全造成巨大威胁的自然灾害。中国地震频发，是全球受地震灾害影响最为严重的国家之一。为了进一步探索地震三要素有效预测这一世界难题，北京大学集成微系统实验室研发了多分量地震监测系统 AETA。该系统从 16 年底至今已稳定运行了 3 年，积累了 TB 级别的原始数据，并从均值、振铃计数、峰值频率和峰值幅度四个方面提取了 GB 级别的特征数据。

目前在现有特征数据方面已经取得了丰富的研究成果，但现有特征数据存在有效信息过少，缺乏数据预处理以及部分特征不够合理的问题。为了充分利用 TB 级别的原始数据，多维度地挖掘特征从而研究信号与地震的相关性，本文搭建了 AETA 原始数据分析平台，主要工作和创新点如下：

1、分析原始数据的特点，并基于差异性大，噪声相对较大和数据量大的特点完成顶层设计。原始数据分析平台由特征生成流和异常风险库两大部分组成，每个模块内部使用生产者消费者模型连接。架构方面可以分为持久化存储层、接入层、数据处理层和展示层。

2、设计与实现了特征生成流。原始数据通过原始数据下载器获取，并在预处理模块识别与处理断电重启异常、数据缺失异常和脉冲型异常。预处理后的数据将在特征提取模块中从时域、频域和小波变换方面提取特征。其中，电磁扰动信号共 49 个特征，地声信号共 42 个特征。特征数据最终将通过特征数据库接入模块存入特征数据库。

3、设计与实现了异常风险库。特征数据在异常检测模块检测得到异常风险指标，并在异常评价模块中通过方差分析，AUC 指标和时间叠加分析三个方面评价异常和地震的相关性。检测到的异常风险和相关性指标将持久化存储在异常风险库中。

4、优化运行时间长的函数并实现加速。在特征提取模块中，短时能量和短时过零率特征运行时间较长。优化后，提取短时能量特征运行时间分别比遍历法和矩阵法减少 87.87%~89.24%和 27.31%~68.41%，提取短时过零率特征比遍历法减少 99.82%。通过优化，单进程提取单台站单分量一天的特征运行时间减少至 17.83s，满足实际需求。

5、基于开源的 python 可视化框架 Dash 实现可视化。数据分析人员可以通过网页端从波形图、数据分布图、频谱热力图以及时空图多角度查看特征数据以及其异常和分布。

关键词：AETA，分析平台，特征提取，异常检测

Design and Implementation of AETA Raw Data Analysis Platform

Binyan Ma (Microelectronics and Solid-State Electronics)
Directed by Xin'an Wang

ABSTRACT

Earthquakes, as known as a serious natural disaster, threading the safety of human life and property. With the frequent occurrence of earthquakes, China is regarded as one of the countries most affected by earthquake disasters in the world. In order to effectively predict the three elements of earthquakes, which is a world problem, the multi-component seismic monitoring system AETA has been developed by the Key Laboratory of Integrated Microsystems of Peking University Shenzhen Graduate School. This system has been operating steadily for 3 years since the end of 2016, accumulating TB-level raw data, and extracting GB-level feature data from four aspects: mean value, ringing count, peak frequency and peak amplitude.

Up to now, abundant research results have been achieved in the existing feature data. But these features have three problems, including insufficient effective information, lack of data pre-processing, and improper method in some situation. In order to make full use of TB-level raw data and mine features multi dimensionally to study the correlation between signals and earthquakes, AETA Raw Data Analysis Platform is designed in this paper. The main work and innovations are as follows:

1. Analyzed the characteristics of the raw data, and completed the top-level design based on the characteristics of significant difference, moderate noise and large amount of data. This platform consists of two parts: feature generation flow and abnormal risk database. Each module is connected internally with the consumer producer model. In terms of architecture, it can be divided into four layers: persistent storage layer, access layer, data processing layer and presentation layer.

2. Designed and implemented feature generation flow. The raw data are obtained through the raw data downloader, and then the dirty data, including instrument restart abnormality, data missing abnormality and pulse type abnormality, are cleaned by the data preprocessing module. After that, features are extracted from time domain, frequency domain and wavelet transform.

Specifically, 49 features are extracted from the electromagnetic disturbance signal, and 42 from geo-acoustics signal. The feature data will eventually be stored in the feature database through the feature database access module.

3. Designed and implemented anomalous risk database. The anomaly detection module detects feature data to get anomaly risk indicators. Then, the correlation between the anomaly and the earthquake is evaluated through three aspects in the anomaly evaluation module. Methods include variance analysis, AUC value and SEA analysis. The detected abnormal risks and correlation indicators will be permanently stored in the abnormal risk database.

4. Optimized long-running time functions for acceleration. Short-term energy and short-term zero-crossing rate features consume a lot of time in the feature extraction module. After optimization, the running time of extracting short-term energy features is reduced by 87.87%~89.24% and 27.31%~68.41% respectively compared with the ergodic method and matrix method. Extracting short-term zero-crossing rate features reduces 99.82% of the time compared with the traversal method. Through optimization, the time required for single feature extraction process to extract features of one station and one component per day is reduced to 17.83s, which meets the actual needs.

5. Visualization based on the open-source Python visualization framework Dash. Data analysts can view the characteristic data and anomalies distribution from the waveform picture, data distribution map, spectrum heat map and space-time map in the web page.

KEY WORDS: AETA, Analysis Platform, Feature Extraction, Abnormal Detection

目录

第一章　引言

1.1　课题背景及研究意义

1.1.1　背景概述

地震作为一种较为常见的自然灾害，给人们的生命和财产安全带来严重威胁， 在全世界自然灾害造成的死亡人数中，54%由地震导致[1]。处于环太平洋地震带与地中海——喜马拉雅山地震带这两大地震带之间的中国是全球受地震灾害影响最大的国家之一[2,3]。自 20 世纪以来，全球 7 级以上强震中，约大约 35%发生在我国。仅在过去 10 年，国内就发生了 2010 年青海玉树 7.1 级地震以及 2017 年九寨沟 7.0 级地震[4,5]，分别造成 2698 人和 25 人遇难。

地震在带来严重的社会危害的同时也激起地震研究的热情，信息科学的不断进步为解决地震预测问题带来新的希望。自 20 世纪 60 年代以来，各种测震和地震前兆数据监测仪器被发明[6-9]。通过这些仪器，人类积累了丰富的数据，现代地震学也随之取得了较大的进展。

对前兆信号的研究是现代地震学的一个研究方向[10]，其中地震的监测和预报研究很早就成为学术界的关注重点。地震预报指在地震发生之前，准确地预测出地震三要素，具体包括地震的时间、地点和震级[11]。准确的三要素信息，对帮助政府决策，进行人、财物的转移和疏散工作具有重大帮助，进一步地达到减少地震造成的人员以及财产损失的目的。

地震预报的准确性依赖于长期有效的地震监测工作的支撑。地震监测需要选择与地震活动相关的特定信号量，使用设备通过信息化的手段进行观测，在长期监测中研究震前，发震时和震后这些特定信号的变化。在此之中，震前信号异常，也称为前兆信号异常，在地震监测中意义最大。对于前兆信号异常，国内外学术界进行了大量研究，包括地应力[12-15]，电磁扰动[16-18]，空间电离扰动[19-21]，地下流体异常[22-24]等等。

北京大学深圳研究生院集成微系统实验室研发了多分量地震监测系统 AETA[11]，通过测量电磁扰动信号和地声信号进行地震三要素预测。从 16 年底至今，系统已经稳定运行 3 年，在全国 12 个省份累计安装设备 200 余套，积累了 TB 级别的原始数据，并从均值、振铃计数、峰值频率和峰值幅度四个方面提取了 GB 级别的特征数据，在地震预报中取得一定成果。但是，现有的特征数据对原始数据的挖掘不够充分，需要进一步从原始数据挖掘信息，以提高地震预报的准确性。

1.1.2 AETA 系统概述

地震是地壳运动的一种表现，它的孕育和发生伴随着地下能量强烈的活动过程。在这个活动过程中会出现应力，重力，电磁力，声音等物理上的异常，以及气体浓度，离子浓度等化学上的异常。不同的观测量需要采取对应的研究方法，观测量的选择是系统设计中关键的一部分。

在地震的前兆研究中，地震电磁学是一个重要的分支。电磁扰动（电场、磁场变化）与其他研究手段，如地质学、地球物理学和地球化学相比，是一种较好捕捉的地震短临异常。在临地震研究中，电磁扰动被视为其中一种最灵敏的前兆信号，往往在在震前数日甚至数小时内出现明显的变化，而且多个案例表明异常幅度与震级正相关[25]。

另一方面，作为一种能直接获取地下信息的途径，地声具有很大的研究价值[26-28]。地震的孕育过程常伴随着地应力场的变化、地壳形变的慢速变化及蠕变，当局部地壳受压而发生形变或微破裂时，将发出地声。此外，在岩石破裂过程中，因高温或其他因素产生的地下电离气体，在接近地面处放电时也会产生地声。

因此，AETA 系统分别使用电磁和地声探头观测电磁扰动信号和地声信号。电磁探头能在 0.1Hz~10kHz 较宽的动态范围内对 0.1nT~1000nT 的甚低频、超低频电磁波段以 500Hz 的采样率和 16 位精度进行监测，灵敏度在此波段满足>20mV/nT。地声探头则涵盖了 0.1Hz~50kHz 的次声波、声波到部分超声波波段，以 500Hz 的采样率和 18 位精度进行监测，灵敏度在此波段满足>20mV/nT。

除了使用电磁和地声传感探头采集电磁扰动和地声信号之外，AETA 系统还包括数据处理终端，监测数据云平台，云平台提取的特征数据将存储在数据库并通过网页端和客户端展示 [5]，如图 1.1 所示。

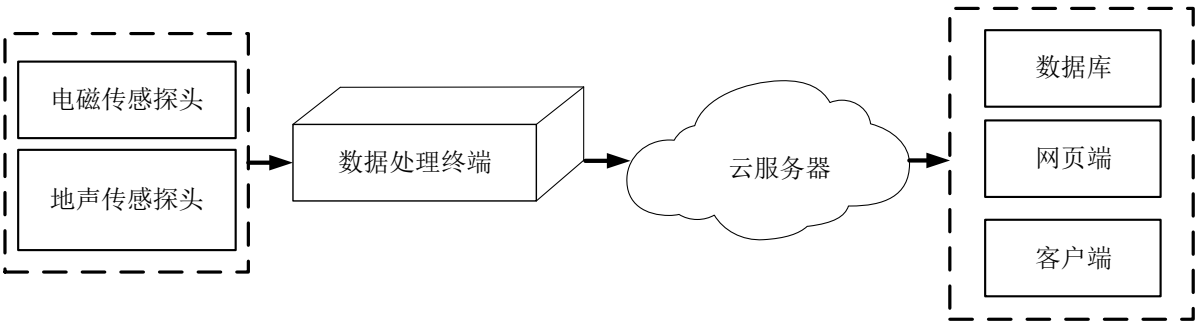


图 1.1 多分量地震监测系统 AETA

2015 年 8 月，AETA 系统的第一版设备的小批量试制顺利完成。2016 年 6 月，第二版设备小批量生产完成，共 20 套。在中国地震局监测预报司和全国各地地震局的支持以及协助配合下，两版设备均进行了现场布设试验。结果证明，对当地震例，AETA 系统具有较好的映震效果，初步验证了系统灵敏性、稳定性和一致性。2016 年底 AETA 项目与专业硬件服务商深圳卓翼科技达成深度合作，完成了第三版设备的开发和生产，

AETA 系统正式步入定型批量生产阶段。2019 年 3 月，AETA 系统从抽样版本升级为连续版本，以更全面地监控和地震相关的前兆信息。截止目前为止，在全国范围内 AETA 系统安装数量超过 200 套，范围遍布河北、四川、云南、西藏、广东和台湾地区，具体分布见图 1.2。

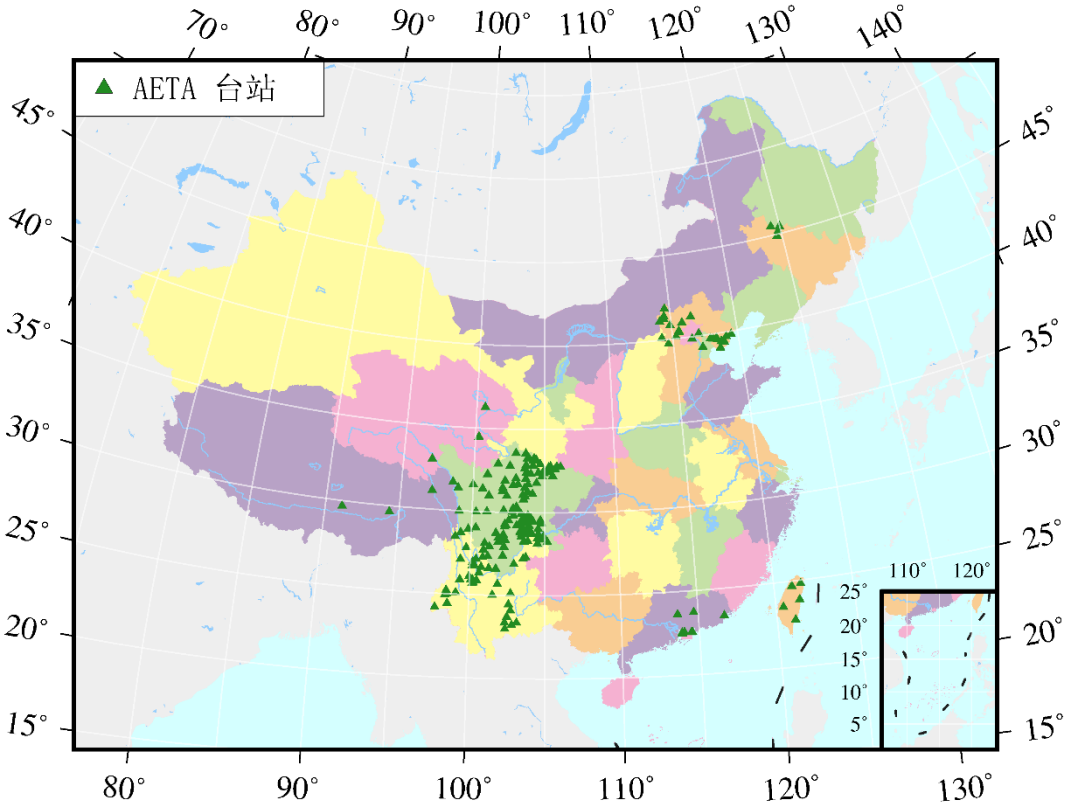


图 1.2　全国范围内 AETA 台站分布图

1.1.3　特征数据的成果和不足

对于采集到的电磁和地声数据，服务器以 1 分钟的颗粒度对原始数据提取特征，并存入阿里云的数据库中。现有的特征数据从均值、峰值频率、峰值幅度和振铃计数这四个角度描述原始数据。其中，均值是时域上的特征，峰值频率、峰值幅度和振铃计数这三个是频域上的特征。

基于现有的特征数据，分析研究发现采集到的观测量具有映震效果。例如，在电磁方面，不少台站出现与日出日落基本同步的 SRSS 波，日出时由高变低，日落时从低变高[29]。多个震例表明，震前 SRSS 波往往会发生出现，消失或变异的情况。对 SRSS 波采用主成分分析（PCA）分解并重构，得到的特征值和具有明显的映震关系[30]，如图 1.3 展示了九寨沟 7.0 级地震前，九寨沟防震减灾局出现的条带异常情况。此外，对于特征数据，特别是均值数据，采用人工免疫算法，ARIMA 算法检测到的异常和地震也有一定的相关性[31,32]。在模型方面，对 2017 年 6 月 1 日至 2019 年 3 月 1 日间，发生

的 1111 次发生在中国境内及周边的地震事件进行数据集切分并用随机森林进行训练和预测，其中 361 例弱震，661 例有感地震，83 例中强震，6 例强震，AUC 达 0.696。

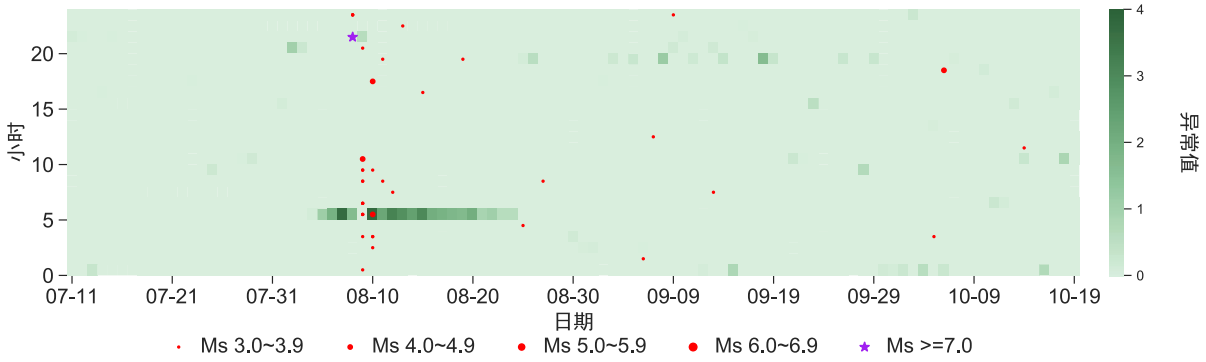


图 1.3　九寨沟 7.0 级地震九寨沟防震减灾局 PCA 算法捕捉的条带异常

然而，现有的特征数据存在以下三点不足。第一，有效信息过少。每分钟 60KB 的原始数据只用 4 个特征量来表示，虽然能展示原始数据中最显著的特点，但丢失数据中更为丰富的信息。地震前兆有用的信息不一定体现在信号总体模式的变化，而在一些细节的改变。特别地，在时域上只有均值这一个特征，只展现了信号在这一分钟整体幅值的大小，缺乏对细节的刻画。第二，缺乏数据预处理，虽然数据处理终端进行了简单的预处理，但未针对以断电重启为代表的异常导致的脏数据进行清洗。第三，部分特征数据设置不够合理。以振铃计数为例，该特征描述的是信号 1 秒钟平均穿过 0 轴的次数。由于采用单一固定阈值，现有的振铃计数特征鲁棒性差，对零点的设定以及噪声较为敏感。由于设备存在零飘现象，再加上背景噪声的影响，计算出的振铃计数不能有效反映原有物理意义，而且缺乏稳定性。

另一方面，地震是一个复杂的地壳活动过程，地震前兆信号可能以不同的方式来体现，现在难以找到任何一种确定的前兆信号，依靠这种高度压缩，丢失大部分信息且稀少的特征值来预测地震三要素，很难达到较好的预测效果，而且对不同地区，不同时间发生的地震也存在缺乏泛化能力的问题。

为了进一步挖掘观测量与地震之间的相关性，更好地实现对地震三要素的预测，从原始数据出发进一步提取特征是必要的选择。考虑到目前 AETA 系统已经积累了 TB 级别的原始数据，提取特征的代价较大，搭建科学的特征生成流完成提取流程具有巨大意义。结合之前的数据分析经验，AETA 各台站间存在较大差异性。和单纯分析特征来预测地震相比，更有效的方式是关注台站自身的变化，提取特征数据中的异常，并结合多台站联合分析。因此需要检测特征数据中的异常事件，设计异常风险库，并进行对应的可视化工作来直观地展示。

1.2　国内外研究状况

在国内，地震行业在近十年的发展后，已建成的观测网络达到巨大的规模，无论是前兆还是测震学科相关的观测数据，每天新产生和累计的量都是海量的规模。因此，使用数据分析平台，进一步提高大数据的数据挖掘能力，进而进行数据分析工作成为迫切需要。

2016年，刘高川[33]等人设计了地震前兆台网数据跟踪分析平台，该平台实现了事件分析提取与录入、产品自动产出等功能，拥有齐全的功能，方便地使用以及好的扩展性三大优点。时间分析记录和专题报告每天能大量产出，从而用于地震分析预报会商和管理台站运维。目前该平台已在全国地震前兆台网大范围部署。

2017年，彭俊芳[34]等人构建了广西地震工程分析信息系统平台属性数据库。系统按不同的专题内容详细设计了8个数据库，并基于ADO技术实现了应用程序与数据库间的接口。数据库表按照相互切换的标签页对话框进行设计，并能够对成果数据进行动态存储和管理。

2019年，康凯[35]等人使用Hadoop架构构建地震大数据平台。该平台具有数据采集、数据重构、数据服务这三个核心业务模块，以及九个功能子模块。此外，针对地震行业数据的分布式存储、分布式检索及数据可视化等功能进行了探索性的实现，并优化了部署方式及可靠性。

2019年，燕云[36]等人设计了辽宁省地震前兆应急监控与数据处理平台。整体架构包括展示层、业务层和持久层，软件方面包括C/S模式开发的辽宁地震前兆台网运维监控系统，和基于B/S模式开发的网络版。

在国外方面，与地震有关的数据分析平台较少，但在与之类似的地理信息平台上面有不少研究。

1999年，NIKOLAOU A.S.[37]设计了一款地理图形信息系统，该系统用户友好性良好，具有交互式图形用户界面和联机帮助窗口。通过该系统能方便地进行概率相关的分析，如确定和显示预期地震运动和破坏的空间分布。

2007年，T. C. Vance[38]等人设计了一个地理信息系统与海洋学和决策支持模型集成的系统GeoModeler。该系统使用基于Java的应用编程接口和连接器，将地理信息系统与区域海洋建模系统和海啸模型分割方法直接连接起来。使用者可以很方便地选择数据设置模型的参数并运行模型。

2011年，Hardisty[39]等人设计了一款开源的、互联网提供的地理可视化和分析工具包GeoViz Toolkit，不熟悉编程的用户可以很灵活地使用工具包提供的组件动态创建自己的地理可视化和分析组件集，集成地理可视化方法号空间分析方法，并向其他用户

分享。GeoViz Toolkit使用了过去十年在软件工程方面的关键进展，包括对象的自动内省、软件设计模式和方法的反射调用。

2015年，Levente J Klein[40]等人针对地理信息高达PB级别的数据量，以及数据格式多种多样的特点，设计了PAIRS这款可扩展的地理空间数据分析平台。该平台建立在开源大数据软件之上，管理数据的自动下载、数据迁移和可扩展存储。通过使用该平台时空数据的自动更新，连表和归一化的功能，用户能快速进行数据挖掘。

当前，国内外学者的地震或地理的数据分析平台主要针对已经提取好且存储的数据，而对于AETA的原始数据，除了需要将特征数据进行分析并可视化外，还需要预先从PB级别的原始数据提取特征，并持久化存储特征数据。本文针对实际数据分析的需求，构建AETA原始数据分析平台，同时注重稳定性和运行速度。

1.3 论文组织架构

本文以 AETA 原始数据分析平台的设计与实现为核心，具体的组织结构如下：

第一章为引言部分。首先概述了地震的危害以及地震预测的相关背景，并介绍了为解决地震三要素预测而对研发的多分量地震监测系统 AETA。接下来说明现有的特征数据方面取得的研究成果以及特征数据的不足，提出未来的研究需要从原始数据提取特征，从而进一步引出构建数据分析平台的必要性。在对国内外地震数据分析平台进行调研后，提出结合 AETA 系统的特点与数据分析的需求，研制一款适用 AETA 的原始数据分析平台。

第二章主要对 AETA 原始数据分析平台的进行相关需求分析和实现总体设计。首先系统介绍了 AETA 原始数据产生的数据流，原始数据的特点和常见波形代表的意义。针对原始数据的特点和数据分析的需求，原始数据分析平台由特征生成流和异常风险库两部分组成，各部分需求也被明确和细化。然后按照拟定的设计原则，对平台进行了初步的架构设计。逻辑框架划分为四层，依次是数据持久化存储层、接入层、数据处理层和展示层。

第三章主要介绍了特征生成流具体的设计与实现。针对从原始数据到特征数据入库的过程，生产者消费者模型被用于串接原始数据下载器模块，数据预处理模块，特征数据提取模块和特征数据库接入模块。此外，针对原始数据的特点，本章还理清原始数据预处理的需求和步骤，确定时域相关，频域相关和小波变换相关的特征，对于特征提取模块中时耗高的部分提出优化的算法，还根据持久化存储的需求和特征数据的特点，设计了特征数据库表的结构和表之间的关系并对特征表水平分表。

第四章主要介绍了异常风险库具体的设计与实现。首先整理了常用的异常检测算法以及针对时间序列的异常检测算法，并在异常检测模块集成了项目组使用中发现有较

好效果的低复杂度算法，包括对单特征进行检测的滑动四分位法、 ksigma，以及多特征的集成检测算法孤立森林。检测出的异常风险特征将通过方差分析，AUC 指标和时间叠加分析三个方面评价异常和地震的相关性。本章还设计了持久化储存的异常风险库，详细介绍了数据库接入器各个接口，并使用基于 python 的 web 可视化框架 Dash 构建展示层。可以通过波形图，数据分布图，频谱热力图和时空图多个角度可视化特征数据的异常及分布。

第五章主要对 AETA 原始数据分析平台进行测试和展示。针对特征生成流和异常风险库，本章从功能和性能对整体和子模块进行测试，最终展现了展示层可视化的结果。测试表明，AETA 原始数据分析平台各层次设计合理，整体功能完善，达到了顶层设计的要求。

第六章为总结与展望，本章总结全文工作和成果，并对后续工作进行了展望。

第二章　AETA 原始数据分析平台需求分析与总体设计

第一章介绍了地震预测研究现状及 AETA 特征数据上已取得的研究成果，通过分析现有特征的不足从而引出构建原始数据分析平台的重要性，进一步地对现有的地理和地震领域的数据分析平台进行了介绍。本章将首先从电磁扰动和地声原始数据的特点入手，结合数据分析的需求和特点，将数据分析平台划分为特征生成流和异常风险库两部分，并对每个部分进行需求分析和确立需要的子模块，最后从整体的逻辑结构角度总结，完成对原始数据分析平台的顶层设计。

2.1　AETA 原始数据

2.1.1　AETA 数据采集流程

电磁扰动和地声信号分别由电磁探头和地声探头采集后，将使用有线方式连接数据处理终端并通过 TCP/IP 协议传输数据。对于接受的数据，终端通过 buffer 缓存并分成两路，一路不做处理，以全频原始数据保存。此时对应的数据的采样频率地声为 150kHz，电磁扰动为 30kHz。另一路数据将依次通过两级滑窗抽样和 FIR 低通滤波，形成数据率为 500Hz，经过 200Hz 低通滤波器滤波的低频原始数据[41]。得到低频原始数据的滤波和抽样参数具体如表 2.1 所示。

表 2.1　AETA 系统中的特征数据

信号分量	滤波次序	输入采样率 (Hz)	滑窗间隔	输出采样率 (Hz)	低通滤波器截止频率 (Hz)
电磁扰动信号	第一级	30k	10	3k	1k
	第二级	3k	6	500	200
地声信号	第一级	150k	15	10k	3k
	第二级	10k	20	500	200

AETA 系统从 2017 年 6 月稳定运行至今，期间数据处理终端的数据采样间隔进行过两次调整。初始版本所有台站的抽样策略为每 10 分钟抽取 1 分钟原始数据并传至应用服务器。在 2017 年 8 月~10 月进行了第一次调整，将所有台站的抽样策略更改为每 3 分钟抽取 1 分钟原始数据。在后续数据分析的过程中，发现绝大多数与地震相关的信息集中在低频段，为了实时捕捉地震前兆信号，2018 年 11 月至 2019 年 3 月进行了第

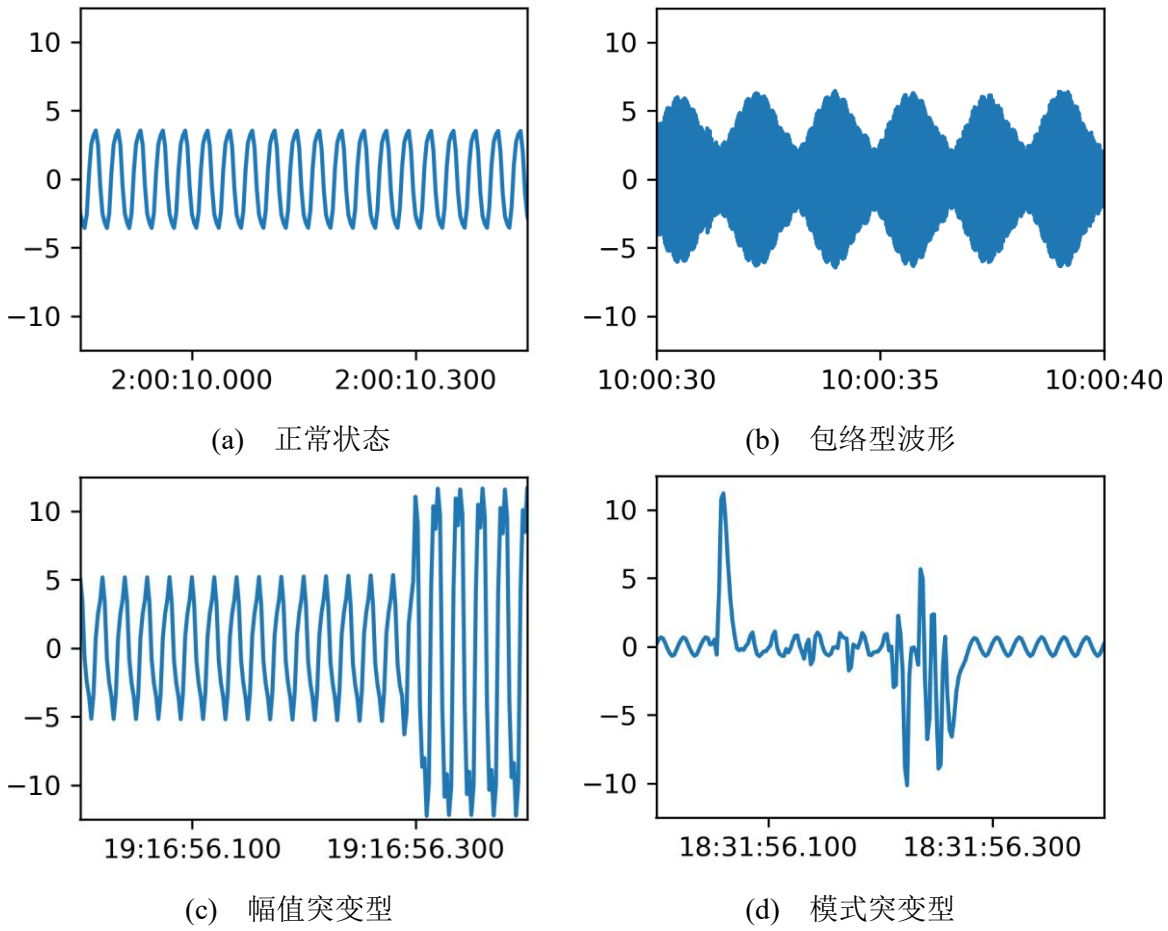
二次调整，取消对时间的窗口取样，并只保留低频原始数据。由于全频原始数据已不再采集，无法通过全频原始数据进行地震预测，本文研究的原始数据特指低频原始数据。

数据处理终端将处理后的数据上传到应用服务器，应用服务器将分支为两部分，其中一部分不作处理，直接以原始数据的形式存入阿里云服务器，最多保存 30 天。另一部分从原始数据中提取四种基础特征值，分别为均值、振铃计数和峰值频率和峰值频率幅度，以特征数据的形式存入云端 MySQL 数据库。

2.1.2　AETA 原始数据特点

电磁扰动和地声的原始数据均以 500Hz 的频率对数据进行采样，通带分别为 0.1Hz~200Hz 和 0.5Hz~200Hz，幅值方面分别为-12.288~12.288V 和 0~5V。除此之外，由于观测量性质的不同，两种有各自对应的特点。首先介绍电磁扰动和地声原始数据常见的波形和蕴含的信息。

(1) 电磁扰动原始数据独有特点



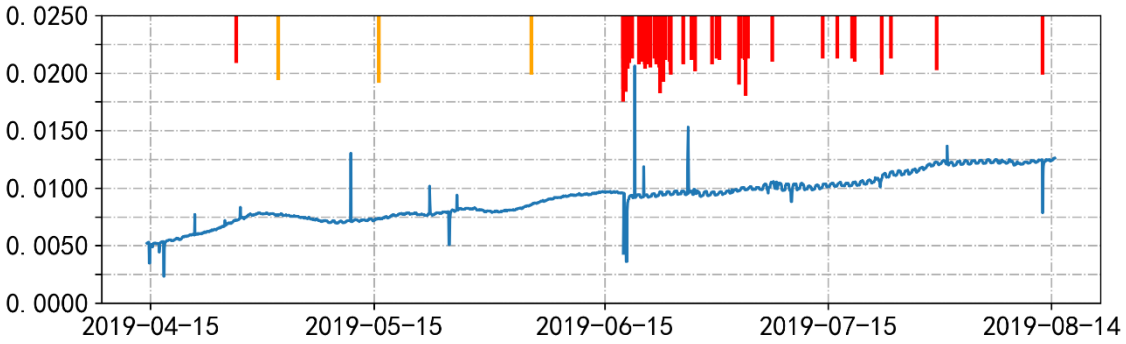
(a)　正常状态　　(b)　包络型波形

(c)　幅值突变型　　(d)　模式突变型

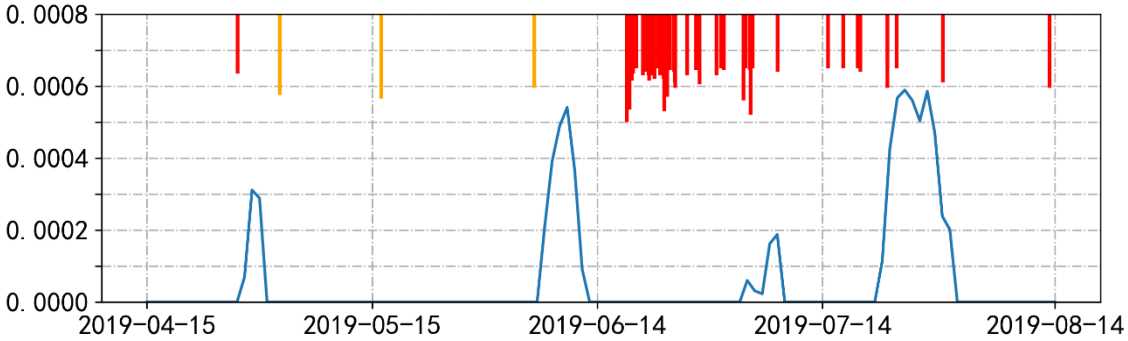
图 2.1　电磁扰动原始信号典型波形

电磁探头检测的是空间电磁场变化的情况，时域上体现了较强的周期性，频谱上

能量绝大部分能量集中在 50Hz 和 150Hz 附近，其中以 50Hz 为主导，整体数据具有较强的冗余性。除了这两个频段的能量，在 30Hz 以下的超低频段（ULF）具有映震关系。图 2.1 分别展示了电磁扰动原始信号常见的波形以及异常，包括包络型，幅值突变型，模式突变型。图 2.2 则展现了 48 号珙县气象台超低频段电磁扰动信号的均值以及按天聚合后滑动四分位法提取的异常值，可以看到存在明显的映震关系。



(a)　超低频电磁扰动信号均值特征



(b)　滑动四分位法异常值

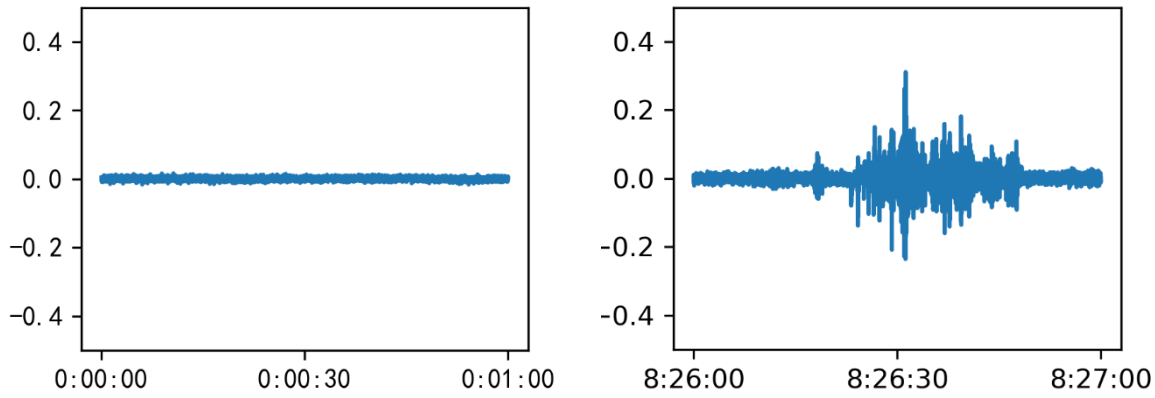
图 2.2　珙县气象局超低频段电磁扰动信号和提取的异常值

根据电磁扰动原始数据的特点，可以发现现有的特征数据存在明显不足。在频域方面，峰值频率特征的取值基本为 50Hz 或 150Hz，可以视为一组二值数据，信息熵过小，提供的有效信息不足。同时峰值幅度和均值具有极高相关度，以 105 号峨眉山防震减灾局台站为例，从 2019 年 1 月 1 日到 2019 年 12 月 31 日，提取的 452871 条特征数据中，两者的皮尔森相关系数达到 0.9803，这两个特征存在冗余。对于其他频段，特别是超低频段，现有的 4 个特征均难以有效体现其内在信息。

(2) 地声原始数据独有特点

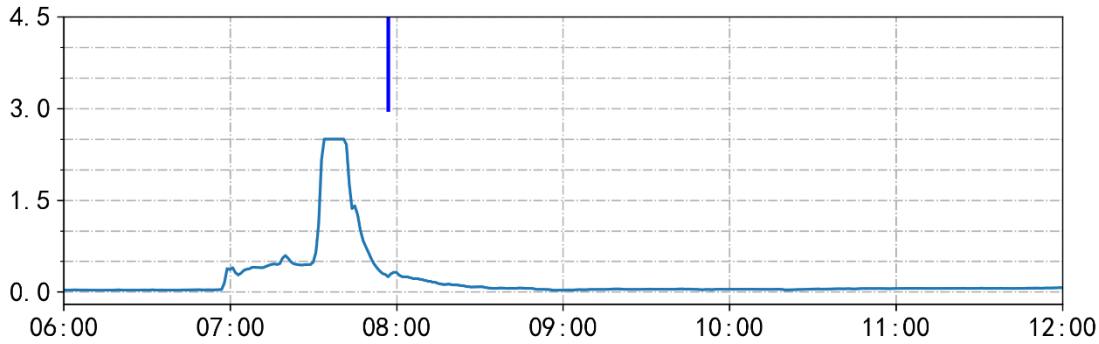
地声探头检测的是地下的振动信号。在正常情况下，探头只采集到微弱的地脉动，信号基本是一条直线。当地面震动时，探头能检测到与声波类似的信号，如图 2.3 所示。对于地声原始信号，值得关注的有效信息主要集中在两方面，分别为突然产生的振动

以及整体均值的漂移。在地震的前后，可能会出现明显的异常，如图 2.4 展示了大鹏新区海啸台在 2019 年 11 月 26 日台湾海峡 5.2 级前后的地声现有均值数据和对应的原始数据。在震前，地声信号出现整体向上漂移再缓慢恢复的过程。

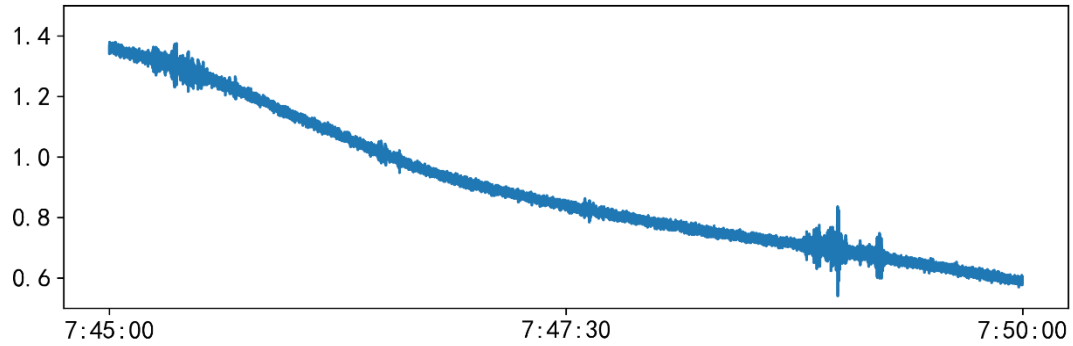


(a)　平静时　　(b)　振动时

图 2.3　地声原始信号典型波形



(a)　地声现有均值特征



(b)　7:45~7:50 的原始数据

图 2.4　2018 年 11 月 26 日海啸台地声信号

另外，地声探头也能灵敏地捕捉到地震波信号，图 2.5 展示了长宁 6.0 级地震在地震发生时，距离震中 188.9km 的峨边中学台站的地声信号原始波形。地震于 22:55:43 发生，发震后 32 秒后捕捉到地震的 p 波，55 秒后捕捉到地震的 s 波。

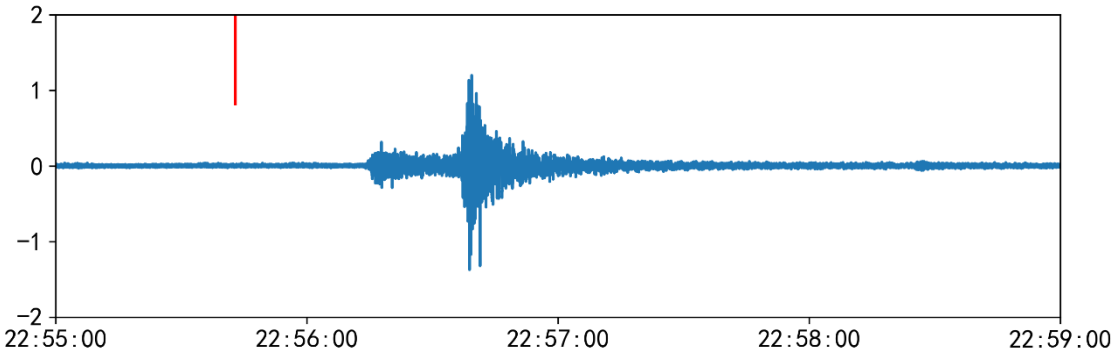


图 2.5　长宁地震峨边中学台站捕捉到地震波信号

地声信号平时振幅极小，现有均值和振铃计数特征对调零值极为敏感。目前地声探头采用固定调零值，在设备的长期运行中，由于温漂等原因往往发生零飘，导致现有特征数据产生误差，严重时甚至完全失真，例如偏差过大时振铃计数会长期处于 0 值。

(3) 电磁扰动和地声原始数据共有特点

对于用于提取特征的原始数据，具有以下三个共有特点。

第一，原始数据存在差异性。这个差异性体现在时空这两方面。首先，对于同一时刻的不同地区的台站，由于经纬度和安装环境的差异，采集到的观测量存在差异。另一方面，同一台站在使用过程中设备性能会缓慢变化，同时温度等背景环境在长期上有所改变，同一台站在不同时刻的数据也可能存在差异，对于相同的特征量，表示的含义也不尽相同。如图 2.6 展示了位于104.25°E, 33.26°N的九寨沟防震减灾局台站和位于101.01°E, 30.03°N的甘孜雅江台站在 2019 年 7 月 1 日的电磁扰动均值特征数据，同一时刻的不同台站电磁扰动均值特征数据模式存在较大差异。同一台站不同时期的数的差异如上文展示的图 2.2，珙县气象局电磁扰动信号在超低频段存在长期缓慢波动。

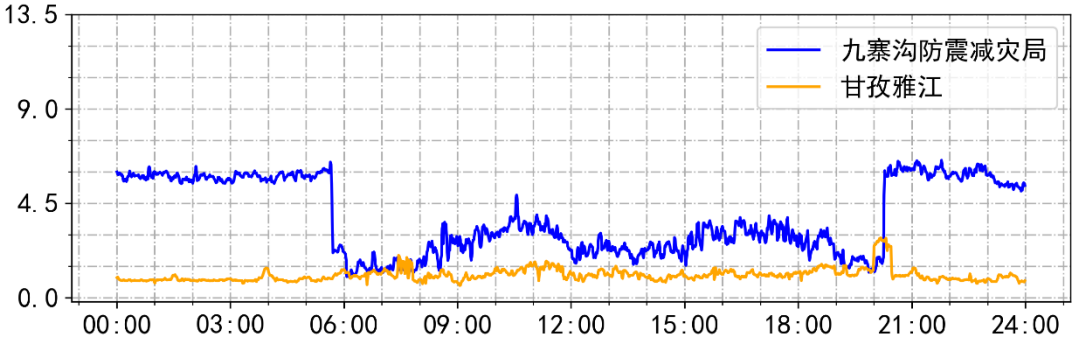


图 2.6　九寨沟防震减灾局台站和甘孜雅江电磁扰动均值数据对比

第二，原始数据噪声相对较大。相比于现有的特征数据，原始信号在提供了更多的细节的同时，噪声相对也相对更大。地震发生前后引起的异常可能较为微弱，往往间接影响观测量。除了地球活动引起观测量的变化外，可能还有由其他因素引起的噪声。

表 2.2　低频原始数据数据量

时间	时间抽样策略	数据传输条数	数据量
2017.06～2017.10	1 条/10min	144 条/天	16.88MB/天
2017.08～2019.03	1 条/3min	480 条/天	56.25MB/天
2018.11～ 至今	1 条/1min	1440 条/天	168.75MB/天

第三，原始数据的数据量大。具体体现在历史数据量大以及数据新增速度快两方面。每条电磁扰动和地声低频原始数据均为 30000 个数据点，每个数据点大小为 2 字节，总体大小均为 60KB。在目前最新的版本下，一个台站一天将分别上传 1440 条电磁扰动和地声的原始数据，每天新增的数据量如表 2.2 所示。按每月 30 天计算，一个台站一天的原始数据为 168.8MB，一个月为 4.95GB，一个月为 60.17GB。目前原始数据文件存储系统已经存储约 20TB 的原始数据，且每天以约 25GB 的速度增加。无论是现有的原始数据量，还是原始数据的增速均较大，因此从原始数据批量挖掘特征时，需要进行合适的规划，尽量一次完成提取流程。

2.2　AETA 原始数据分析平台的需求分析

在之前的小节中，介绍了 AETA 系统从电磁扰动和地声两个分量探测与地震相关的前兆信号，并在提取的特征数据取得了一定成果。由于信息量过小，缺乏数据预处理和部分特征不够合理，未能有效挖掘局原始数据中的信息，需要从原始数据出发挖掘。原始数据在时域和频域上具有丰富的信息，同时具有时空上具有差异性，其噪声相对较大且历史数据量大三个共有特点。

鉴于现有特征未能很好反映原始数据中蕴含的信息，为了重新挖掘原始数据，同时考虑到巨量的历史数据的生成的代价大，需要构建特征生成流来高速有效地提取特征。

由于原始数据在具有时空差异性，同样的特征数据在不同台站，或同一台站不同的时间意义不一定一致，但在短时间跨度内同一台站的数据具有可比性，研究应更关注某一个台站和自身相比的异常。进一步考虑到采集到的信号具有较大的噪声，相比单个台站的异常，更合理的方式是联合多个台站进行分析。因此，需要构建异常风险库。

综上分析，AETA 原始数据分析平台将由特征生成流和异常风险库这两个部分构成，针对每个组成部分，下面将需求拆分并进行详细分析。

2.2.1　AETA 原始数据特征生成流的需求分析

本小节对特征生成流的需求进行拆分，从原始数据文件存储系统获取数据到将特征数据持续化存储与特征数据库的过程进行介绍。

1、计算历史积累原始数据的特征值，并每日自动提取台站新产生并上传原始数据的特征值。在 AETA 系统平稳运行的 3 年多以来，已经积累且需要挖掘的历史数据约为 20TB。对于这一部分数据，数据量庞大，计算成本高，需要以数据流的方式连贯完成整个提取流程。同时，需要确保每天采集的原始数据能及时提取特征，为未来实时预报做准备。

2、获取需要计算特征值的原始数据列表。具体可以分成已储存在原始数据文件系统的历史数据和每日新上传的数据这两种情况。其中后者包括台站当天上传的数据，以及补传过去的数据。对于历史数据，可以通过 http 接口按照台站号和数据对应的时间戳逐渐获取。对于后者，则通过按台站号和文件存储日期的 PostgreSQL 接口来获取。

3、根据原始数据列表获取对应的原始数据，并进行数据转换。通过 http 协议或数据库接口获取指定台站，指定分量和指定区间内的原始数据。为了减少数据传输量，获取的原始数据为 16 位的整型数据，需要根据相应探头调零和放大参数将其转化为对应的电压。

4、对原始数据进行数据预处理。设备运行和数据传输的过程会对采集到的原始数据引入非自然因素引起的脏数据，需要对数据进行合适的清理工作，以提高后续提取特征的质量。

5、多维度对原始数据提取特征。地震的前兆信号可能在不同维度体现在电磁扰动和地声的原始信号上，需要从时域、频域和小波变换等多个维度选择特征，多维度地挖掘原始数据的信息。这部分特征不仅在异常风险库中使用，也可以作为新的基础特征，进行更多的挖掘，用于算法研究。

6、设计合适的特征数据库，利用数据库接入器的接口实现特征数据的持久化存储。根据拟提取的特征数据的种类和数据量，设计合理的表结构和表间关系，以快速地实现表特征数据的插入和读取，同时编写合适的 API，完成对表的操作。

7、记录运行日志。在合适的位置打点，记录特征生成流的进度和每个环节运行的状态。

2.2.2 AETA 原始数据异常风险库的需求分析

从原始数据在时域、频域和小波变换多个角度提取特征并存入特征数据库后，需要提取特征数据出现的异常，总体需求如下。

1、 获取特征数据，统一不同时间段数据的时间颗粒度，以方便后面的环节使用异常检测算法检测特征数据的异常。

2、 检测特征数据的异常。鉴于原始数据的时空的差异性将反映在提取的特征数据上，相比于简单的特征数据，更值得关注的是单个台站某个特征数据的异常表现。

3、 选择合适异常评价指标，衡量异常和地震之间的相关性。提取出的异常风险指

标可能反应了地震前兆信号，也可能由其他因素导致。因此需要有一个合适的评价指标来衡量某种捕捉到的异常和地震相关性。

4、 设计合适的异常风险特征库来储存异常指标。结合特征数据和异常种类的特点，设计合理的异常风险库表结构和表间关系，同时编写合适的 API，以方便其他环节插入或获取异常风险。

5、 可视化数据及异常指标。为了更好的进行数据分析，需要对原始数据，特征数据和异常事件从波形图，数据分布图，频谱热力图，时空图等多个方面实现可视化，直观地展示数据的分布和变化，从数据层面上加深引起异常机理的理解。

6、 记录运行日志。作用和特征生成流的相同。

2.3　AETA 原始数据分析平台的框架设计

根据 2.2.1 的需求，特征生成流的框架如图 2.7 所示。特征数据生成流将由原始数据下载器，数据预处理模块，特征提取模块，特征数据库接入模块以及监听模块组成。对于涉及数据操作的模块，模块间将通过生产者消费之模型来进行信息传递。在提升系统整体运行性能的同时实现模块间的解耦，提高可维护性。具体的实现方式通过共享内存缓冲区完成。另外，被监听的模块在每次获取或完成一项任务时将向共享变量写入心跳信息，即当前时间和任务状态。监听模块将监控各个线程的运行状态，若发现有线程心跳异常，将重新启动该线程，同时将记录该线程未完成的任务到错误日志中，方便获取未成功完成的任务详细信息，排查对应错误。

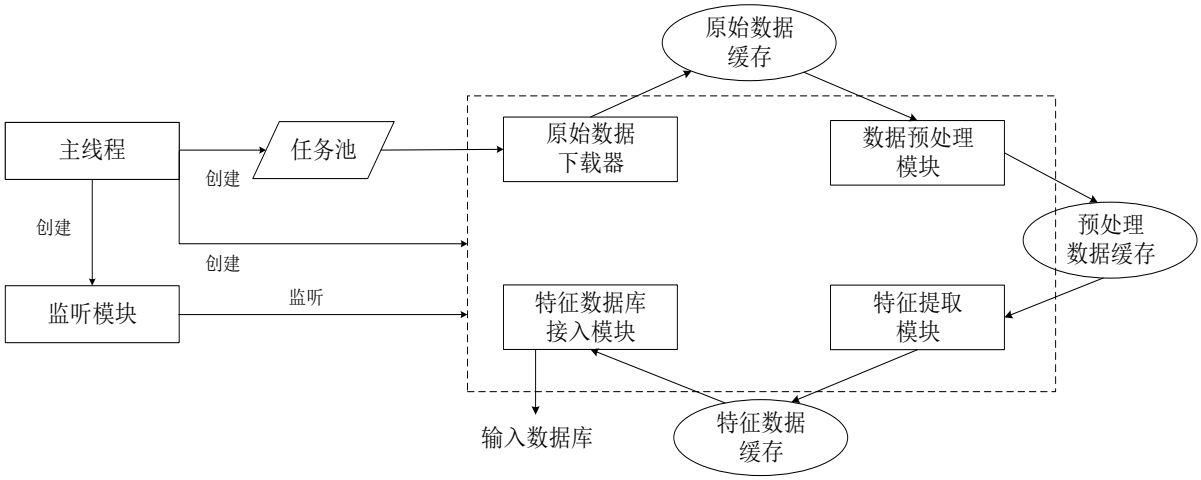


图 2.7　特征生成流框架图

结合图 2.7，　特征生成流运行过程如下：

1、 主线程启动后，启动各个线程并创建任务池，数据缓存区和共享变量区。然后根据需要生成特征的台站列表，信号分量以及时间区间拆分任务，生成任务列表，并放

入任务池。每个任务为一个台站，一个分量一天的原始数据。

2、 原始数据下载器从任务池中获取任务，并进一步获取对应的原始数据文件信息列表。根据原始数据文件信息列表下载原始数据，解码并转换为探头采集的电压后，将转换后的数据放入原始数据缓存中。

3、 数据预处理模块从原始缓存中获取原始数据，使用算法判断并去除因设备或传输导致的异常，按照探头类型和调零放大配置进行数据转换，最后将预处理后的数据放入预处理数据缓存中。

4、 特征提取模块从预处理数据缓存中获取预处理数据，对每一分钟的原始数据从多个角度提取时域、频域和小波变换相关的特征。将提取的特征数据进行汇总后，放入特征数据缓存中。

5、 特征数据库接入模块从特征数据缓存中获取特征数据，利用数据库接入器提供的接口与 MySQL 数据库进行通信，将特征数据入库。

6、 监听线程监听 2~5 的所有线程，对于存在心跳异常的线程，获取线程处理数据对应的任务信息并记录对应错误日志中，最后重新启动存在异常的线程。

根据 2.2.1 的需求，异常风险库的框架如图 2.8 所示。其原理和总体流程和特征生成流相似。异常风险指标在历史提取完成后会每天只新增当日异常，而异常评价模块将每隔较长时间才会更新一次。

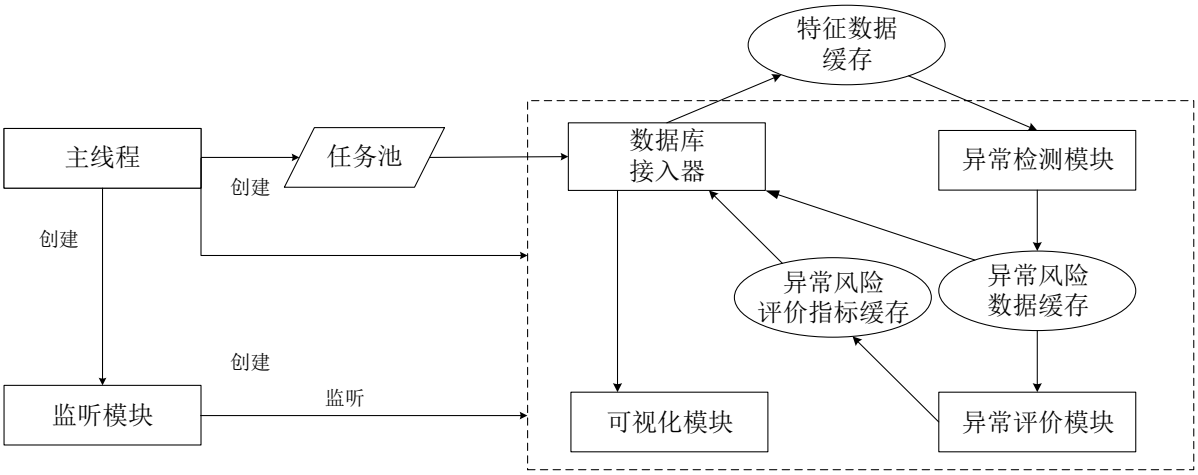


图 2.8　异常风险库框架图

2.4　AETA 原始数据分析平台逻辑架构

将特征数据流和异常风险库整合，可以提取出如图 2.9 所示的 AETA 原始数据分析平台的整体逻辑架构。该系统可以划分为持久化存储层、接入层、数据处理层和展示层共四个层次。下面将以自底向上的方式依次介绍各层次。

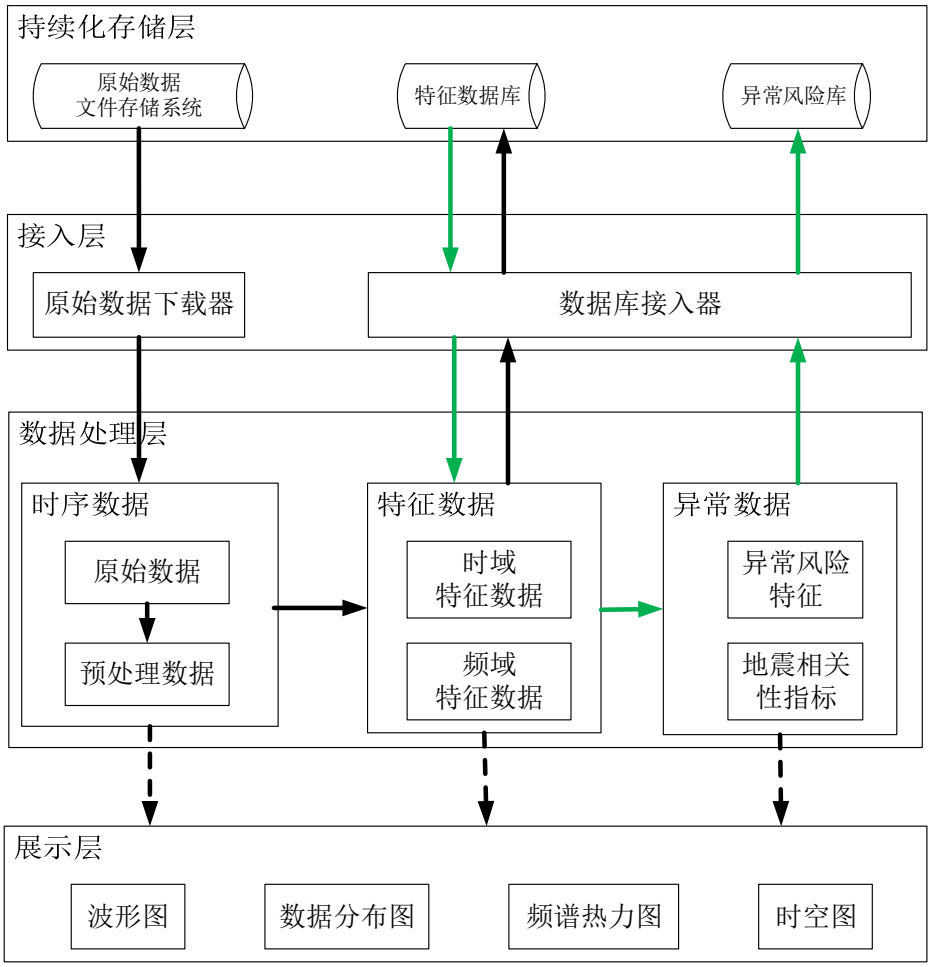


图 2.9　AETA 原始数据分析平台整体逻辑框架

1、持久化存储层

持久化存储层为 AETA 原始数据分析平台中的各项数据提供持久化存储支持，这里的存储技术主要是数据库。持久化存储层中存储的数据主要有三类，第一类是原始数据，通过 MongoDB 文档类数据库存储；第二类是从原始数据新挖掘的特征数据，按照台站号和信号类型分表，最后一类是提取到的异常数据。后两者存储在 MySQL 关系型数据库。

2、接入层

接入层是数据处理层和持久化存储层的中间代理，通过原始数据下载器和数据库接入器完成对原始数据文件存储系统以及特征数据库，异常风险库的交互，同时接受与送达数据处理层的数据。接入层可以避免数据处理层对底层数据的直接操作，提高数据安全性。

3、数据处理层

数据处理层对接入层的数据进行处理，包括原始数据的预处理，预处理数据的特

征提取和对特征数据提取异常风险特征，并判断异常风险特征和地震的相关性。原始数据，预处理数据，特征数据和异常数据将在该层进行传递和转换，最终生成的数据将返回接入层或传递到展示层用于可视化。

4、展示层

展示层对数据处理层的数据进行可视化展示。展示层通过波形图，数据分布图，频谱热力图和时空图分别将数据在时间上的特性，频谱上的特性以及数据在大范围时空下的分布展示出来，有助于从数据层面上加深引起异常机理的理解，同时给多台站联合分析，判断地震三要素提供帮助。

2.5 本章小结

本章首先介绍了 AETA 数据的采集流程，讲述了原始数据如何从探头采集到最终存储在原始数据文件存储系统的流程，并详细讲述了电磁扰动和地声数据的基本形态，同时原始数据具有差异性大，噪声相对较大和数据量大的三个特点。基于原始数据的特点，AETA 原始数据分析平台将由特征生成流和异常风险库这两大部分组成。接下来，本章对这两部分的需求进行了分析，进一步地设计了整体的结构。最后，本章介绍了 AETA 原始数据分析平台的总体架构，自底而上依次为持久化存储层、接入层、数据处理层和展示层共四层。

第三章　AETA 原始数据特征生成流的设计与实现

本章将介绍特征生成流部分的各个组成模块的框图和细节，讲述存储在原始数据文件储存系统的原始数据从下载，预处理到提取时域相关的特征、频域相关的特征和小波变换相关的特征，最后持久化存储在特征数据库的全过程。各个模块间将通过生产者消费者模型进行连接，为了提高特征提取速度，针对运行时间长的短时能量，短时过零率函数进行算法上的优化。为了提高数据库的性能，对特征数据表进行水平分表处理。

3.1　生产者消费者模型

生产者消费者模型为软件设计中常用的一种设计方法，其示意图如图 3.1 所示。

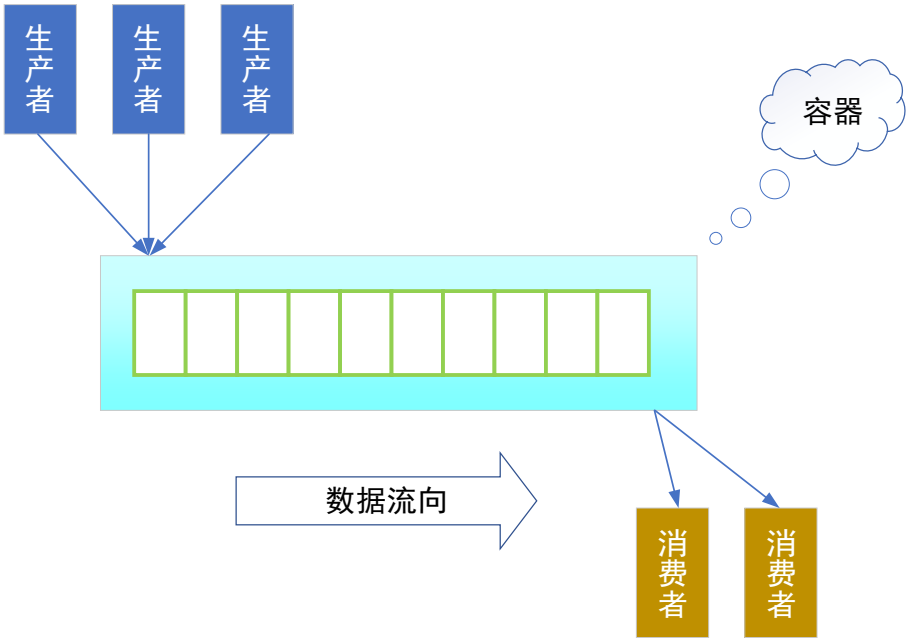


图 3.1　生产者消费者模型

生产者消费者模型关注的是线程模型中的经典问题，生产者和消费者通过共享的存储空间连接。产品由生产者源源不断地生产并放入到储存空间中，最后由消费者取出。当空间满时将触发消费者阻塞。反之，当空间满时，生产者将被阻塞。在整个模型过程中，生产者消费者模型中存在三种关系。在生产者与生产者之间存在互斥关系，消费者与消费者之间存在互斥关系，而在生产者与消费者之间，即存在互斥关系，也存在同步关系。

生产者消费者模型能平衡生产者的生产能力和消费者的消费能力，同时减少查询次数，降低 CPU 的使用率，从而提升整个系统的运行效率。另一方面，生产者消费者模型能实现生产者和消费者之间的解耦，降低生产者模块和消费者模块的联系，从而降低代码复杂度，提高程序的可维护性。

此外，本文采用 python 来实现原始数据分析平台。Python 是解释型语言，存在全局解释器锁 GIL，在运行的时候 python 会锁定解释器。在 Cpython 解释器中，线程执行 CPU 指令需要 2 个条件：

1. 被操作系统调度出来（操作系统允许它占用 CPU）
2. 获取到 GIL（Cpython 解释器允许它执行指令）

这导致 python 在运行代码的时候，总体流程为：设置 GIL -> 切换到一个线程去执行 -> 运行 -> 把线程设置为睡眠状态 >切换到一个线程去执行>…>结束。在每个时间点只有一个线程在运行。所以，python 多线程只是伪多线程，对于计算密集型任务，python 的多线程不能提高速度，应采用多进程，加快计算速度。对于 IO 密集型任务，CPU 运行时间短，大量开销在 IO 及等待时间，可以直接用多线程。

在 AETA 原始数据分析平台中，各个模块间将通过生产者消费者模型进行管理，生产者和消费者之间将通过数据缓冲区隔离。对于 IO 密集型的模块，生产者或消费者将用多线程完成，如数据库接入器。对于计算密集型模块，将采用多进程完成，如数据预处理模块和特征提取模块。

3.2 原始数据下载器的设计与实现

数据处理终端接收到探头采集的信号后，每分钟将生成一条大小为 60KB 的*.data 文件，并以采集时间戳作为文件名。服务器接收这些原始数据，并存入原始数据文件存储系统中。对于大量的这种小文件，文件存储系统使用 GlusterFS 分布式文件系统存储，并用 MongoDB 记录并管理每个文件的信息。对于存储的原始数据，它提供了 http 和 PostgreSQL 两个数据接口下载指定台站，指定分量在特定时间内的原始数据文件。原始数据下载器针对任务池中的下载任务，从原始数据文件储存系统中下载数据，并将转换后的数据传到原始数据缓存中。对于每个下载线程，具体的框架如图 3.2 所示。

由于在原始数据文件中只有采集到的信号分量的数值，没有台站，时间和信号分量的信息，在进行原始数据下载之前，需要提前获取原始文件的信息，包括对应的信号分量，时间戳和文件大小。有两种接口可以选择，http 接口可以通过台站号、信号分量和数据对应的时间戳获取，而使用 PostgreSQL 开发的接口可通过台站号、信号分量和存储的时间来获取。对于历史数据可以采用 http 接口获取，而对于当日上传的新数据，由于存在补传，http 接口会遗漏对应的数据，应该选用 PostgreSQL 接口来获取。

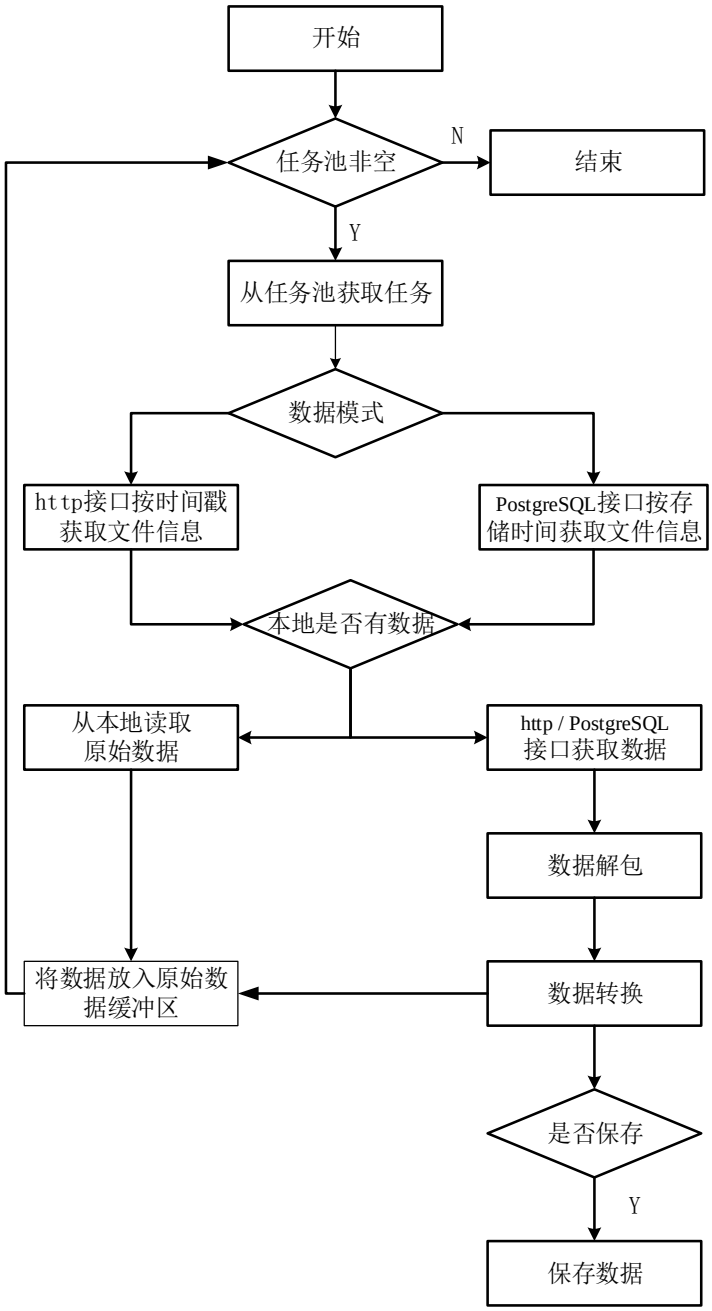


图 3.2　原始数据下载器模块执行框图

除了在特征数据流外，日常的分析中也需要使用原始数据下载器，直接针对原始数据进行更多的探究性工作。如果都通过网络传输获取数据，将都到网速限制，因此在原始数据下载器还添加了本地存储的读取功能。当设置优先从本地读取数据时，如果本地存在数据，则跳过下载步骤，直接从本地读取并将数据放入原始数据缓冲区。

当设置重新下载数据或者本地未找到数据时，将通过 http 接口或 PostgreSQL 接口下载数据，并根据原始文件的信息将每条数据拆分。下载的数据不直接为探头采集到的电压，而是 16 位大端编码的整型数据。因此需要先将二进制解码为证书，然后再将

数据根据探头类型转换成探头实际采集到的电压数据，转换公式如下：

$$V = x \cdot V_{max}/32767 \tag{3.1}$$

其中V_{max}为探头 ADC 的量程。对于电磁扰动探头，V_{max}为 12.288V，对于地声探头，V_{max}为 5.000V。

转换结束后的数据将放入原始数据缓冲区中，随后开始新的循环，从任务池中获取下一条任务。如若需要保存，将开启一个保存进程，按照参数设置的文件格式，将原始数据保存至本地。

原始数据下载器通过封装成一个类来实现，可以选择下载的台站，信号类型，日期区间，每日时段等参数，具体的参数如表 3.1 所示。

表 3.1　原始数据下载器参数列表

参数名	含义	数据类型	默认值
stationID	台站列表	list / tuple	[]
signalType	信号类型	list / tuple	[1] (表示电磁扰动数据)
timeRange	下载时间范围	list / tuple	过去一周
dailyRange	每天时间范围	list / tuple	[“0:00”, “24:00”]
replace	是否重新下载	bool	False
save	是否保存到本地	bool	True
shareDict	共享内存中的字典	manager.dict	None
downloadMode	下载方式	http 或 db	http
savePath	保存路径	str	./
returnType	返回数据格式	data,csv,pkl 之一	pkl
threadNum	线程个数	int	2
compress	是否压缩	Bool	True
log_level	显示日志级别	str	“info”

3.3　数据预处理模块的设计与实现

数据预处理指的是在特征挖掘等后续工作之前，对数据进行前置处理，从而提高数据的质量。在实际的环境中，数据往往存在噪声干扰，缺失值以及数据格式不一致的情况，直接对脏数据提取特征将影响后续数据挖掘的效果和所需的时间。因此，致力于提升数据质量的数据预处理模块是特征生成流的必要组成部分。数据预处理有多种方法：

数据清理，数据填补，数据变换等。在工程实践中，数据预处理没有万能的方法，也没有标准的流程，通常需要根据任务和数据集的特性灵活处理。具体到 AETA 采集的原始数据，数据预处理模块主要需要进行数据转换和数据清理。

对于传感器连续采集的波形数据，异常区间和离群点具有重大的研究意义。然而，部分异常为设备或网络通信导致，这些与地震前兆无关的异常会增加后续数据分析的难度，需要预先进行处理或标记。针对 AETA 采集的原始数据，需要处理的异常共有三种，分别为断电重启异常、数据缺失异常和脉冲型异常，这些异常需要分别识别并修复或标记。此外，地声信号还需用带通滤波器去除 50Hz,150Hz 附近信号。数据预处理模块的整体框架如图 3.3 所示。

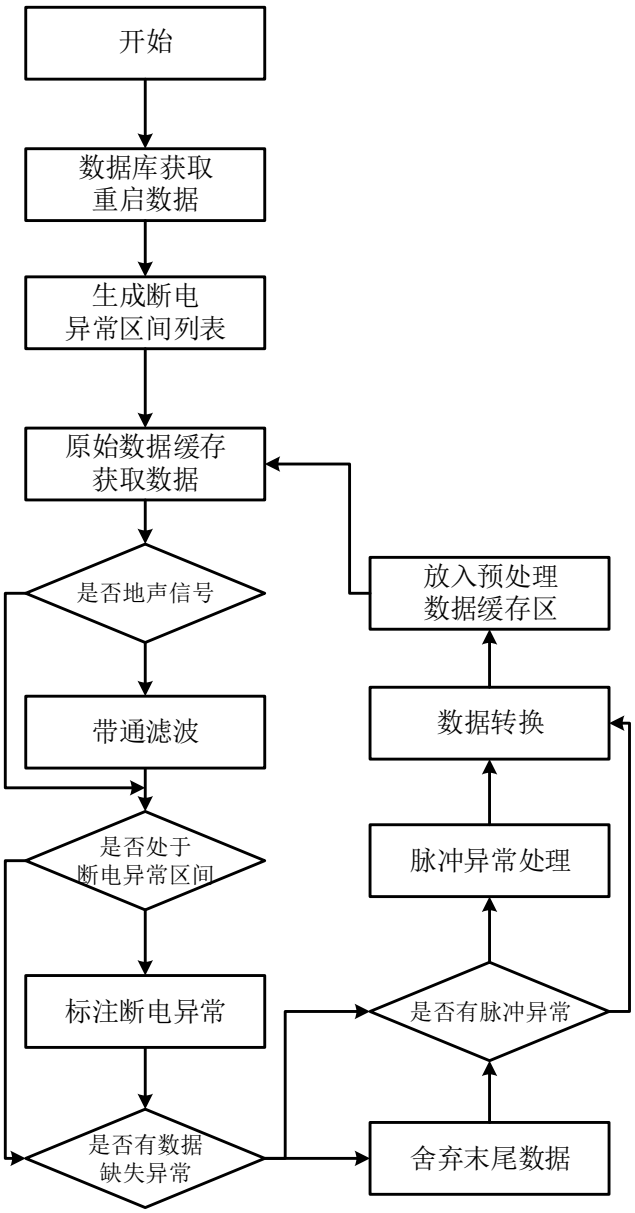
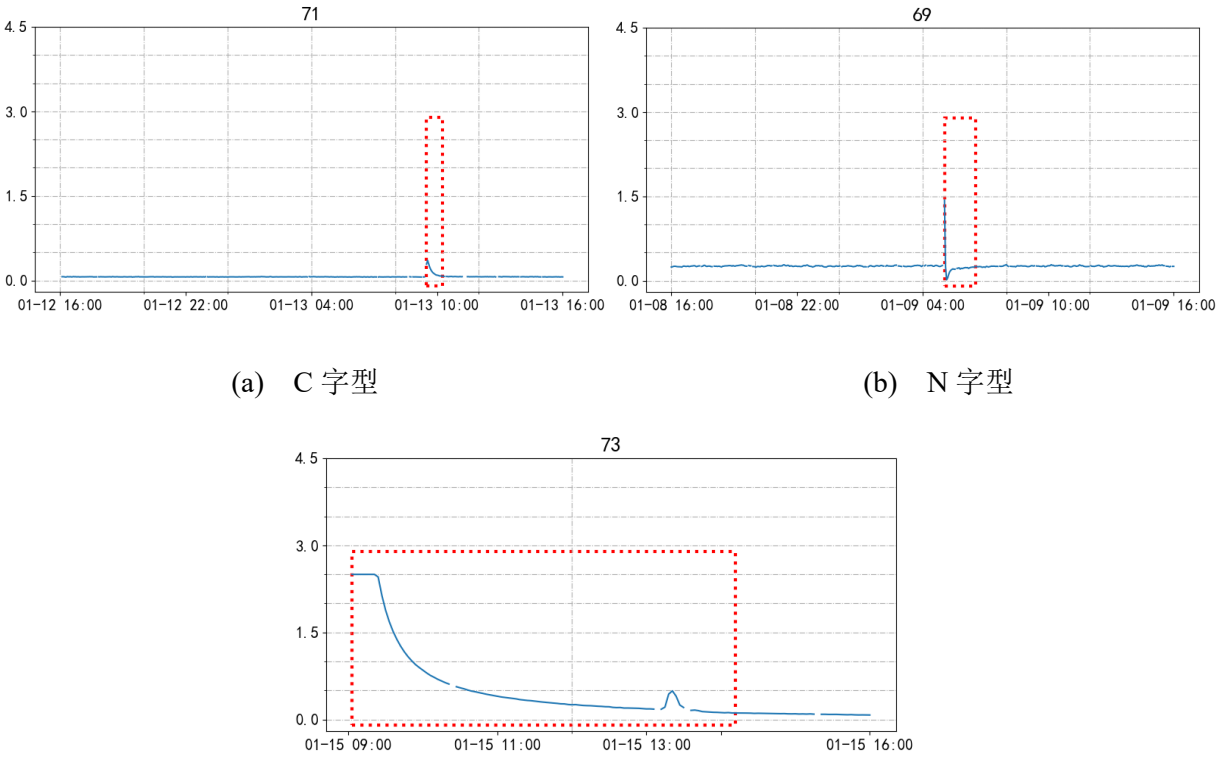


图 3.3　数据预处理模块框图

下面分别仔细讲述各种异常的特点和处理方法

1、断电重启异常

地声探头启动后需要一定时间从环境温度稳定至到工作温度，在早期的版本中，地声探头温度系数大，温漂现象明显。这种异常可以在现有的地声均值数据中明确反映，分别表现出“C 字形”、“N 字形”、“拖尾型”，如图 3.4 所示，其中红框为识别的断电异常区间。



(a) C 字型　　(b) N 字型

(c) 拖尾型

图 3.4 断电重启异常在地声均值特征的体现

断电重启异常在均值特征中能很好地反映，因此只需要采用现有均值数据进行判断。为了检测断电重启导致的异常，首先需要获取每个台站数据处理终端重启的时刻。在云端 MySQL 的 terminallog 表中记录了终端的日志，在终端重启时会发送包含”Terminal system start!”字段的日志，里面包含了对应的重启时间和当前版本信息，因此可以检索对应终端重启的时间和版本号。在实际操作中，每天会有一个定时脚本自动检索终端重启的日志，提取出关键信息并记录在 terminalrestartlog 中，因此可以直接读取 terminalrestartlog 表中获取重启时间。下一步根据断电后的数据是否处于正常的区间范围内，判断是否发生了温漂的现象。具体算法如表 3.2 所示。

表 3.2　提取断电重启异常区间算法

算法　提取断电重启异常区间算法
输入： $D = \{t_d : v_d\}$：台站地声均值数据集 T_r：台站所有重启时间 m：异常判断模式 k：ksigma 原则判断异常的阈值 k l_j：判断正常数据范围的数据长度 输出： $T = \{t_s, t_e\}$：断电造成的异常数据区间 方法： 1. 初始化 $T_s \leftarrow \Phi$。 2. 从T_r获取一个重启时间t，用二分查找找到重启后的第一个数据点t_s。然后往前回溯l_j的时间得到数据区间v_n，作为正常数据。 3. 根据 ksigma 原则从v_n获取正常的数据区间v_T。 4. 选择重启后开始 10 个点，依次判断是否在正常区间外并投票，如果不超过半数，回到第 2 步。 5. 从t_s开始对t_d往后遍历。当v_d回到正常区间v_T且一阶差分符号转变 2 次后，记录结束时间t_d。 6. 将t_s, t_e放入集合T，并回到第二步。 7. 返回断电造成的异常数据区间T。

2、数据缺失异常

正常的电磁扰动和地声原始数据包含 30000 个数据点，每个点大小为 2 字节，总大小为 60kB。然而，有些原始数据存在数据丢失的现象，从某个点开始数据全为 0，如图 3.5 显示了地声探头 ADC 采集到的电压值。该异常集中分布于网络差的台站。设备运行中会发生发送失败的情况，这些没成功上传的原始数据将暂时写入 SD 卡，等待网络环境良好时补传。在长期的运行中，对 SD 卡读写次数超过其寿命，导致数据丢失，读取的都是默认的 0 这个默认值。现有的特征数据在计算时未对这一部分处理，导致电压转换后出现异常值。在某些极端情况下，甚至存在短时间内连续发生的现象，进一步地引起异常区间，这对后面的数据分析引入误导。如图 3.6 展示了乐山防震减灾局 5 月 20 日的地声均值数据，除了 10:21 分外，从 17:26 分到 19:21 的时间段里，间断出现了 27 次数据丢失的现象，产生了异常区间。图 3.5 为其中一个数据缺失异常终端传上来地声探头的电压值。

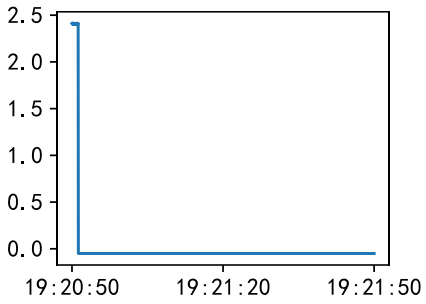


图 3.5　原始数据中数据缺失异常

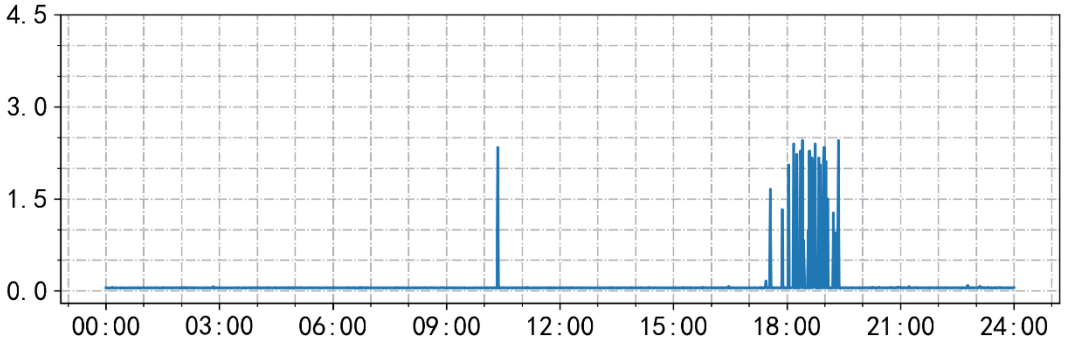
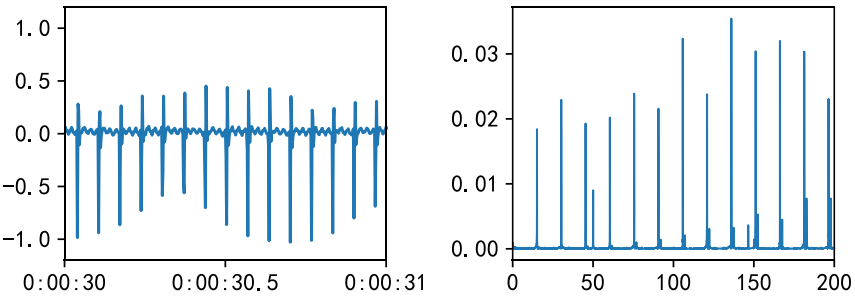


图 3.6　数据缺失异常导致的地声均值异常区间

对于数据缺失异常，可以通过判断原始数据末尾连续若干个数据点是否均为 0 来实现，在实际应用中设为 100 个点。对于检测到出现数据缺失异常的数据，将使用二分查找找出第一个变为 0 的异常点，并用缺失值代替突变点及后面的数据。在接下来的模块中，只会采用前面的数据来提取特征。

3、脉冲型异常

部分台站存在长期或一段时间内持续出现等间隔脉冲异常的现象。脉冲的频率一般为 7~15Hz 之间，且脉冲的高度略有区别。此外，电磁扰动和地声数据往往同时发生频率相同的脉冲。对于这种频率为f_s的脉冲，频谱上也存在间隔f_s的尖峰，如图 3.7 所示。



(a)　时域　　(b)　频域

图 3.7　脉冲型异常

检测算法方面，提供了基于时域和基于频域的两种检测和修复算法。基于时域的算法主要根据信号的一阶差分在脉冲附近明显大于附近的值，通过 IQR 算法判断一阶差分的异常点，并使用开运算连接临近异常点。根据判断异常区间的中点是否大部分等间隔来检测脉冲型异常。检测出的异常区间将通过线性插值填补。具体如表 3.3 所示。

表 3.3　基于时域检测和修复脉冲型算法

算法　基于时域检测和修复脉冲型算法
输入： v_i：AETA 原始数据 T_i：IQR 倍数阈值，默认为 5 T_a：IQR 异常阈值，默认为 1% N_{CRN}：异常区间中点间隔出现前 n 名 T_{CRN}：异常区间中点间隔集中率阈值
输出： v_o：修复后的 AETA 原始数据
方法： 1. 对v_i求一阶差分得到dv_i。 2. 计算dv_i的均值M，上下四分位差$IQR = Q_3 - Q_1$。 3. 计算异常区间上限$U = M + T_i \cdot IQR$。 4. 对dv_i中大于异常区间的点标为 0，其他的标位 1，生成掩膜 mask。若异常点个数小于T_a，直接返回v_i。 5. 对 mask 以大小为 3 的核进行开运算，即先腐蚀再膨胀，从而将间隔过小的异常区间连在一起。 6. 对 mask 差分从而快速找到每个异常区间的起始点，进一步得到异常区间的中点A_m。 7. 对A_m差分得到每个异常区间中点的间隔DA_m，统计出现的DA_m次数C_m。 8. 计算C_m前N_{CRN}的频次占比，得到集中度CRN。 9. 若$CRN > T_{CRN}$，则将v_o中 mask 为 0 的位置设为缺失值，并用线性插值填补得到v_o，否则直接返回数据原时域波形。

基于频域的检测修复算法利用频域上的等间隔性，通过判断超过阈值的频率是否等间隔，从而判断这一分钟的原始数据是否出现过脉冲型异常。首先确定基频f_s的大小，然后将基频和倍频附近的频谱设为 0，最后通过反傅里叶变换转换为时域。注意到大于 250Hz 的频率会有频谱混叠的现象，如图 3.7 的频谱中 100~200Hz 中每个高的尖峰旁边的小尖峰实际为 300~400Hz 对应的频谱。在研究发现，大于 400Hz 倍频的幅值很小，而且原始信号频谱能量主要集中在低频部分，为了不影响低频部分的频谱，只去除不

超过 430Hz 的倍频，具体如表 3.4 所示。

表 3.4　基于时域检测和修复脉冲型算法

算法　基于频域检测和修复脉冲型算法
输入： v_i：AETA 原始数据 T_H：最大百分比阈值，默认为 1% T_e：倍数误差百分比，默认为 5% T_n：集合中元素最小数目，默认为 5 输出： v_o：修复后的 AETA 原始数据 方法： 1. 对v_i进行快速傅里叶变换得到频谱f_i，基于对称性，取前半段f_{ir} 2. 对f_i中除 50,150Hz 附近的以外的值进行排序，得到最大T_H位置的数，作为频谱幅度阈值f_T 3. 找到频谱f_{ir}中大于f_T的所有区间，并以每个区间中最大值f_{max}代表该区间，组成集合F 4. 当F非空时，新建集合D，并将F中的最小值f_0放入 5. 对F中元素f_j从小到大遍历，设F中当前最大值为f_{m}，若$abs(\lvert f_j - f_m \rvert / f_0 - 1) < T_e$，则将$f_j$放入集合$F$ 6. 若F中元素大于T_{n}，对于频谱中$f_j \in (\mathrm{k}f_0 - T_e f_0, kf_0 + T_e f_0)$，小于 250Hz 部分设为 0，250~430Hz 部分将$500 - f_j$设为 0。 7. 重复步骤 4-6，最终得到新频谱f_{ol}，将f_{ol}第一项到倒数第二项翻转并取共轭，得到频谱右半段f_{or}，拼接得到转换后频谱f_o 8. 如果所有集合F个数都小于T_{n}，直接数据原时域波形v_i，否则将f_o进行反傅里叶变换得到v_o并输出

数据清洗的工作完成后，需要进行数据转换。原始数据缓冲区的数据为探头 ADC 采集到的电压值。对于地声数据，电压范围在 0~5V，在安静环境下约为 2.5V。实际使用时需要进行调零和放大操作，具体公式为：

$$V' = (V - V_o) * A \tag{3.2}$$

其中，V_o为调零电压系数，表示安静环境下的电压值，典型值为 2.48V。A为放大系数，放大微弱的振动，典型值为 16。调零放大系数各台站有所差异，需要读取各台站对应配置。

3.4　特征提取模块的设计与实现

3.4.1　时间序列特征提取方法

对于时间序列，一般来说都具有趋势性和周期性的特点。另一方面，对于不同领域的时序数据，会带有独特的领域特点。例如，在金融数据中，具有明显的自相关属性，在趋势上往往呈现“高峰厚尾”的特点；在语音信号中，信号的幅值主要分布在零附近，这是语音信号不持续的特性导致的；对于心电信号，周期性是明显的特征，在频域具有较低的重复率，各波段的频率相对独立[42]。时序数据的特殊性，也在对特征提取在针对性方面提出了更高的要求，提取的特征矢量既要保持原有时间序列的性质，又要凸显不同数据的区别，这样才能确保特征的质量，提高后续训练学习的效果。

常用的时间序列特征提取算法，可以划分为以下四大类：

1、基于基本统计方法，使用时间序列的基本统计量作为特征。具体可以分为时域和频域两大类。在时域上，常见的有均值，方差，极大值，极小值，过零值，高阶统计量等。在频域上，常见的有信号的功率谱，功率谱密度，中值频率等。针对不同类型信号的特点，可以提取对应独特的特征，如在脑电信号(EEG)中，一般提取峰值，熵值[43]，非线性能量[44]，或者针对波形，提取 QT 间隔，ST 间隔，以及 QRS 波各自的峰值[44,45]。

2、基于模型，通过用模型拟合时间序列，然后将模型的系数作为特征值。对于平稳的时间序列，常用的方法有 AR（自回归模型），MA （移动平均模型），将两者结合的 ARMA （自回归移动平均模型）。对于非平稳的时间序列，数据需要反复差分直到得到平稳时间序列，然后再进行拟合，如 ARIMA 模型。特别地，在金融时序数据方面，ARCH 模型(自回归条件异方差族计量模型)能很好地利用金融市场价格围绕估值上限波动的特点[46]。

3、基于变换，通过将时序数据进行变化，突显适合学习的特征，具体可以分为时频变换和线性变换两大类。时频变换将信号从时域变换到频域，最常见的为 FFT（傅里叶变换）[47]，针对 FFT 在时间上缺乏分辨率的问题，还有 STFT（短时傅里叶变换）和 DFT（小波变换）。此外，语音信号上常使用倒谱分析，包括梅尔倒谱和线性预测倒谱[48,49]。在线性变化上，常用 PCA（主成分分析），它用较少相互垂直的分量替代原有高维数据，在起到降维的效果的同时，还能起到去噪的作用。PCA 得到的特征值和特征向量可以选为特征值。此外还有 K-L 变换，奇异值分解等。

4、基于分形维数，使用代表空间扩展程度的分形维数提取特征。分形具有无限精细的结构，在比例上具有自相似性[50]，体现在结构上具有内在规律性。同时它的分数维大于它的拓扑维数，具有复杂性。大自然大部分物体为分形物体，如雪花，海岸线，

叶子。分形维数主要针对非线性信号提取特征，具体可以使用相似维数，盒维数，关系维数等提取作为分形特征。但对于分形维数的算法，其复杂度均较大，因此只适用于少量样本。

3.4.2 AETA 原始数据特征提取

作为传感器采集到的数据，AETA 原始数据一种是典型的时序数据，具有数据量大和单点信息密度低的特点。特别地，对于原始电磁扰动数据，具有明显的周期性，频域上具有单峰或双峰，各波段的频率相对独立，需要更多挖掘超低频段的特征。而原始地声信号与语音信号较为相似，幅值主要分布在零附近，波动较为不连续。同时，除了与前兆相关的信号外，还存在变化时间与人类活动时间相关的信号，这些信号与已知地震波的信号在频谱上有一定区别，也需要从频谱上进一步挖掘特征。

由于原始数据量大，需要选择低时间复杂度，不太依赖预设参数的算法，因此主要采用基于基本统计的方法，同时需要优化算法复杂度，使得特征尽量在$O(n)$，最多在$O(\log\ n)$的时间复杂度内提取。由于无论是电磁扰动原始信号或地声原始信号，都需要在时域外进一步分析，因此需要采用傅里叶变换和小波变换分别提取频域相关的特征和小波变换相关的特征。

此外，为了减少接下来存储在数据库的数据量，提取的特征需要是具有代表性，基本的特征。对于可以通过基本特征内部计算得到的特征，将不被录入。例如标准差可以有方差计算得到，某个频段的能量占比可以由该频段能量大小与总能量相除得到，这些特征将不进行计算。

基于以上特点，原始信号提取的特征分成三大类：时域相关的特征，频域相关的特征，以及小波变换相关的特征。前两者主要从时域和频域出发，提取统计特征，后者采用了基于变换的方法，使用 db4 小波分解重构，能多尺度细化分析信号在时频上的特点。特征数据方面，电磁扰动原始信号一共 49 个特征，地声原始信号一共 42 个特征，具体如表 3.2 所示。

在时域方面，统计特征主要包括方差，极值以及高阶统计量。由于极值对异常点过于敏感，原始信号本身噪声较大，提取了 top5%，top10%对应的值作为特征。鉴于信号可能在 1 分钟的长时间跨度上非平稳，而在短时间跨度上是平稳的，提取短时分析相关的短时能量以及短时过零率作为特征。由于电磁扰动信号需要针对超低频段数据进一步分析，主要能量集中在超低频段的地声信号不需要专门提取超低频段特征，在时域方面电磁扰动的特征数大约为地声的两倍。此外，短时过零率和频率有较大关联，是语音信号处理中的经典特征。由于电磁扰动原始信号主要是 50Hz 或 150Hz，信号具有长时间的固定周期，短时过零率基本无效，因此只针对和语音信号相似的地声信号进行提取。

表 3.2　原始信号特征列表

特征类型	名称	中文含义	电磁特征数	地声特征数
电磁扰动时域 电磁扰动超低频段时域 地声原始时域	var	方差	2	1
	power	功率	2	1
	skew	偏度	2	1
	kurt	峰度	2	1
	abs_max	绝对值的最大值	2	1
	abs_top_x	绝对值最大 x%位置	4	2
	energy_sstd	短时能量标准差	2	1
	energy_smax	短时能量最大值	2	1
	s_zero_rate	短时过零率均值	0	1
	s_zero_rate_max	短时过零率最大值	0	1
电磁扰动频域 地声频域	power_rate_atob	频谱中 a~bHz 功率	11	11
	frequency_center	重心频率	1	1
	mean_square_frequency	均方频率	1	1
	variance_frequency	频率方差	1	1
	frequency_entropy	频谱熵	1	1
电磁扰动 小波变换 地声小波变换	levelx_absmean	第 x 层重构后绝对值的均值	4	4
	levelx_energy	第 x 层重构后能量	4	4
	levelx_energy_svar	第 x 层重构后能量值方差	4	4
	levelx_energy_smax	第 x 层重构后能量值最大值	4	4
合计			49	42

在频域方面，主要通过傅里叶变换得到信号的频谱，提取不同频带的功率值，并从重心频率，均方频率，频率方差以及频谱熵描述频谱。具体的频带为 0~5Hz，5~10Hz，10~15Hz，15~20Hz，20~25Hz，25~30Hz，30~35Hz，35~40Hz，40~60Hz，140~160Hz 以及其他频段，一共 11 个。

在小波变换方面，采用在处理岩石声效果较好的 db4 作为小波基，进行 6 层小波分解[51]。考虑到电磁扰动和地声信号的特点，需要关注超低频段，采用第 4~6 层重构的

细节部分以及第 6 层的近似部分，对应的频率分别为 15.63Hz~31.25Hz，7.81~15.63Hz，3.91~7.81Hz 以及 0~3.91Hz。对提取的波形，再分别提取 4 个统计特征。

特征提取模块的整体框架图如图 3.8 所示。

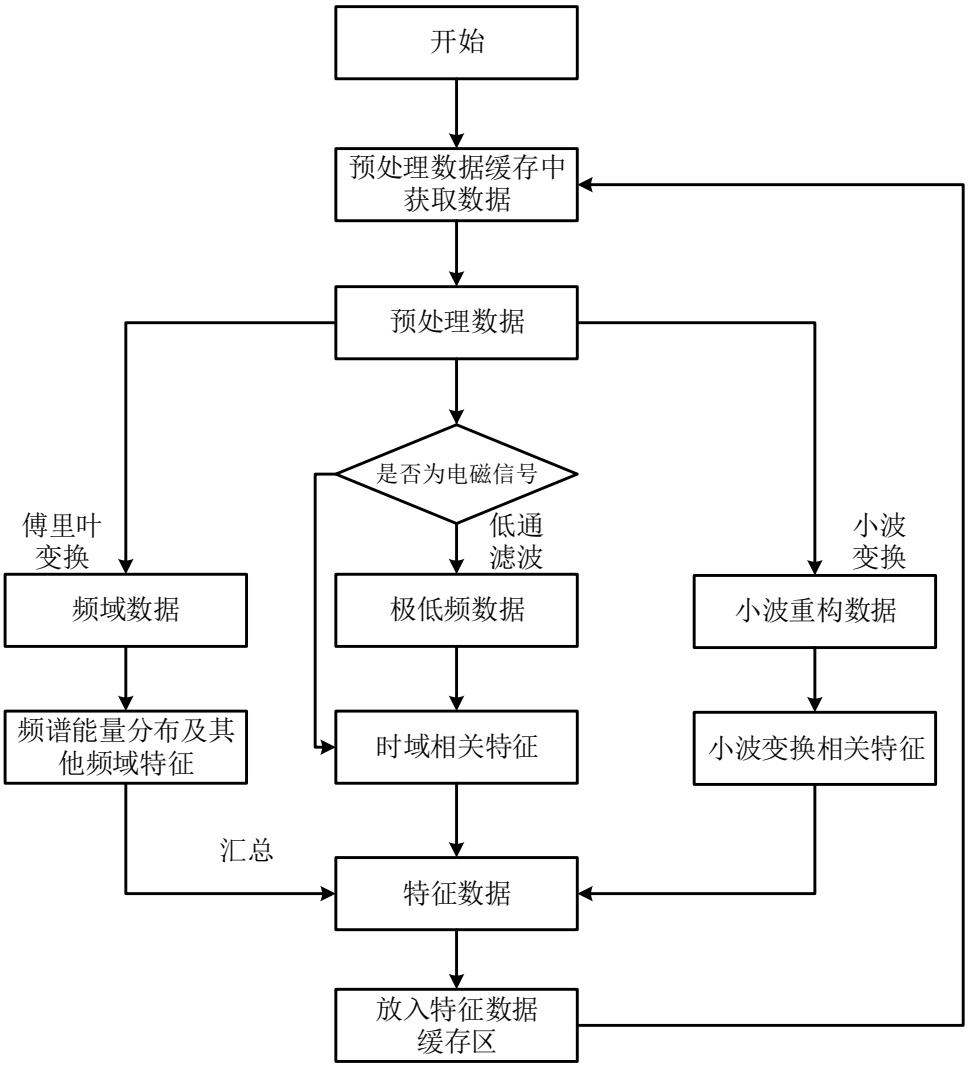


图 3.8　特征提取模块框架图

特征提取模块从预处理数据缓存中获取预处理数据之后，将提取时域统计特征。如果为电磁扰动信号，则进一步地进行低通滤波得到超低频段数据。为了减少计算量，而且电磁扰动信号对相位要求不高，低通滤波器采用阶数少，计算快的 IIR 滤波，为了通带较为平坦，具体选择 6 阶巴特沃斯低通滤波器，截止频率为 30Hz，对应数字频率为 0.12π。接下来分别将预处理数据进行傅里叶变换和小波变换提取频域相关特征和小波变换相关特征，提取的这些特征将根据时间戳作为链接键进行拼接。提取完成后，将特征数据，任务批次信息以及反映特征名及其含义的字典一并放入特征数据缓存区，并获取下一批预处理数据。与之前的模块相同，特征提取模块每个批次处理一个台站，一个特征一天的数据。

3.4.3　算法优化加速

特征提取模块涉及大量的计算，除了使用多进程充分利用 CPU 计算资源来减少运算时间外，需要提高模块内部算法的运算速度，减少运行时间。在特征提取的算法中，耗时最长的是短时相关的特征。在时域特征以及小波变换相关的特征中，需要调用短时能量和短时过零率的函数。对于电磁扰动数据，一共调用 6 次，地声则一共调用 5 次。短时相关的函数未优化前耗时在总运行时长的占用比例中超过 90%，本小节将具体讲述如何优化加速相应的算法。

1、分帧加窗加速

短时相关的特征在语音信号处理中较为常用。对于人的语音信号，或与之类似的 AETA 地声信号，随着时间的推移，其特性及内在特征将逐渐变化，是一个非平稳态过程。但在短的时间内，这些信号可以看成一个准稳态的过程，可以进行分段并分析具体参数，每一段称为一帧[52]。提取短时特征前需要先进行分帧，为了让帧与帧之间平稳过渡，帧与帧之间存在重叠。如图 3.9 展示了总长度为 12，帧长为 4，帧移为 2 时的分帧过程，后文也以这个参数举例说明对应算法。

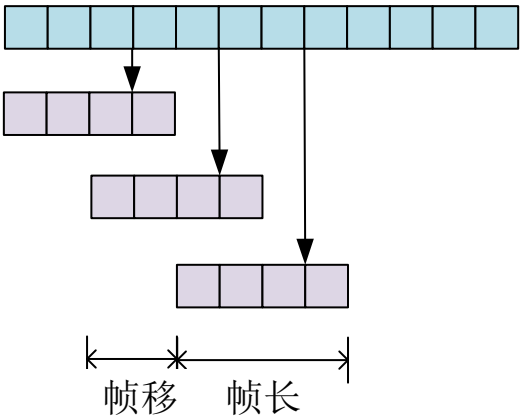


图 3.9　分帧图示

为了尽量满足周期性截断，从减少频谱泄漏的现象，避免出现吉布斯效应，让全局更加连续，在分帧后需要进行加窗操作，将帧内数据和窗函数相乘。本文采用汉明窗加窗，对应公式如下。

$$w(n) = 0.5 + 0.5\cos\left(\frac{2\pi n}{N-1}\right) \tag{3.6}$$

短时相关的特征耗时主要在分帧和加窗上，直接采用遍历的方法对每个分帧加窗效率很低。虽然相比于原生 python，第三方库 numpy 采用 C，部分用 Fortran 实现底层，底层使用 BLAS 做向量、矩阵运算，能利用多核并行计算，这些特性使 numpy 极大地提升速度。但直接使用遍历获取切片，特别是获取某个值 numpy 速度较慢，而且遍历

法无法对每一帧进行并行计算。因此，需要对算法进行优化，优先使用矩阵运算，布尔查找等方法来提高计算速度，并充分发挥并行计算的性能。

针对分帧和加窗，采用矩阵加速计算最简单的方法是直接获取每一帧的数据拼接成矩阵，然后将拼好的分帧数据和窗函数的转置相乘，这样可以充分发挥矩阵并行计算的优势，如图 3.10 所示。特别地，如果需要计算短时能量，需要先将两个矩阵的元素分别平方再运算。

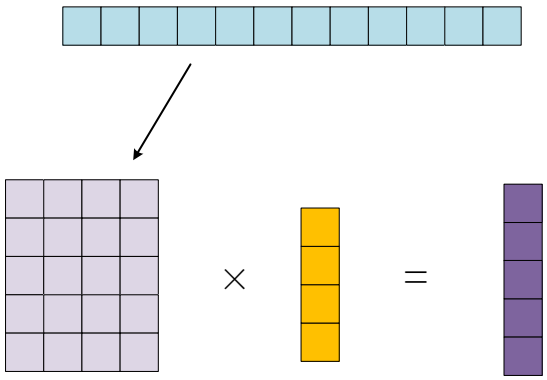


图 3.10　矩阵运算实现分帧加窗

在实现方面，第一帧的位置设置为 0~帧长-1，使用 numpy.tile 重复第一帧合适的次数，然后对第 n 帧加上帧移的 n 倍，即可以生成每一帧数据的位置，从而一次生成所用帧的数据。

直接获取每一帧数据虽然可以用矩阵乘法加速，但需要将原数据的不同部分进行拷贝(copy)。拷贝需要开辟新的空间，运行时间长于和视图(view)相关的操作，而且帧移越小，消耗空间越多。为了避免拷贝，需要尽量使用视图操作。当帧长是帧移的整数倍时，可以直接对原数据的切片并转换后再进行矩阵运算，从而避免拷贝的操作，提升效率，具体流程如图 3.11 所示。

首先以第一个点作为起始点，按帧长平移直到能获得的最大长度作为结束点。通过切片操作选取对应的数据并组成矩阵，使用矩阵乘法得到这一批次的结果。接下来每次将起始点平移帧移的长度，重复之前的操作。具体平移的次数为帧长/帧移。如果某一批次的得到的结果长度小于最大长度，需要在末尾补缺失值。下一步将各个批次的结果拼接起来，并将拼接结果重组得到按分帧顺序每一帧的结果，最后去除末尾的缺失值，只保留有效数据。

设数据总长度为l_d，帧长为l，帧移为d，某一批次起始点为s，则该批次的结束点e为：

$$e = l\lfloor (l_d - s)/l \rfloor + s \tag{3.4}$$

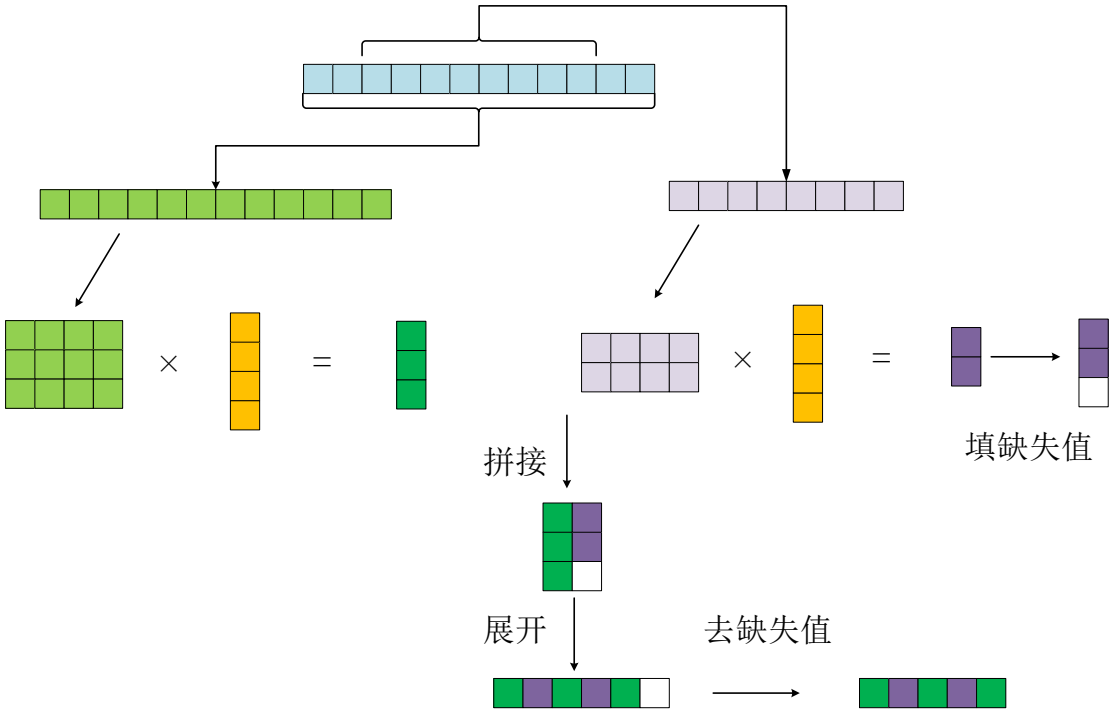


图 3.11　分帧加窗快速算法图示

2、短时过零率加速

短时过零率是一种基本的时域特征，语音信号处理中常与短时能量搭配，使用双阈值进行语音活动检测。短时过零率表示每帧内信号上下穿过零值的次数，能在一定程度上反映信号的频率，具体表达式为：

$$Z_n = \left|\mathrm{sgn}\big(x(n)\big) - \mathrm{sgn}(x(n-1)\right| \tag{3.5}$$

其中，sgn为符号函数，即：

$$\mathrm{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{3.6}$$

对于单门限短时过零率，使用 np.sum((data[1:]*data[:-1])<0)即可得到结果。

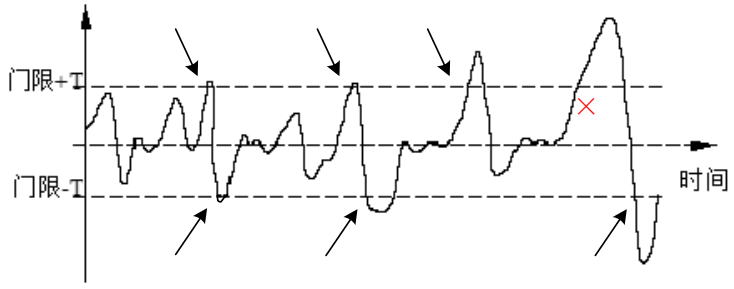


图 3.12　双门限短时过零率

采用单门限计算过零率对噪声过于敏感，噪声使没有信号的时候电压也会在 0 上下波动，造成得到的过零率失真。本文借鉴斯密特触发器，使用双门限机制计算短时过

零率。原理如图 3.12 所示，当穿过门限和上一次反向时过零数加一，如箭头部分所指的位置。对比而言，如果“x”指向的位置穿过门限和上一次同向，则无需增加过零数。

双门限阈值不能像单门限判断相邻数据符号是否变化进行计算。采用遍历算法，即通过一个标识符记录当前的状态，遍历得到结果，对于 1 天 1440 条数据的耗时将大幅增加到约 3 分钟左右。相比而言更高效的方法是避开 numpy 元素遍历的坑，使用 numpy 中的矢量运算，具体如表 3.5 所示。

表 3.5　基于时域检测和修复脉冲型算法

算法　短时过零率快速算法
输入： X：某一帧数据 T_H：双门限的阈值 输出： r：短时过零率 方法： 1. 查X中满足$-T_H \sim T_H$的位置，保存为 mask 数组。 2. 对于处于$-T_H \sim T_H$之间的数据，根据 mask 找到之前第一个不在$-T_H \sim T_H$的数据。 3. 将X中缺失值部分向前填充，使双门限退化成单门限。 4. 使用单门限计算短时过零率 r。 5. 如果某一帧开头是 np.nan，短时过零率 r 加 1。 6. 返回r。

3、其他加速

其他进行加速的手段，包括以下 4 点，主要是避免使用拷贝，多采用视图。

1、展平矩阵时，用 np.ravel()代替 np.flatten()

2、使用 out 参数，如 a = a+1 改为 np.add(a, out=1)

3、能用 numpy 的地方不用 pandas，能用 pandas 的地方不用 list

4、合适的地方使用 numba 加速

3.5　特征数据库及接入模块的设计与实现

对于每一批次的任务，在特征提取模块生成特征并放入特征数据缓存后，将通过特征数据库接入模块调用数据库接入器接口，将特征数据保存到数据库。由于后续异常风险库接入模块也需调用数据库接入器接口，数据库接入器的设计与实现将在第四章进行介绍。

在设计特征数据库接入模块之前，首先需要设计对应的特征数据库。特征数据库主要用于储存特征数据。此外，台站升级会对采集的数据造成影响，最直接的是抽样时间间隔从 10 分钟变为 3 分钟再变为 1 分钟，而且不同台站的升级的时间不一致，因此特征数据库也储存了每个台站版本变化的时间。

AETA 原始数据量大，特征提取模块生成的特征数据的条数也是个庞大的数字。对于每个台站，在当前版本下每天电磁扰动和地声将分别生成 1440 行特征数据，按 200 个台站的规模计算，每天将新增 57.6 万行特征数据，每月新增 1728 万行数据，每年新增 2.1 亿行特征数据。如此多行的数据如果存储在单个表中，将大量增加 I/O 操作次数，降低插入和查询的速度，所以需要通过分表来降低单表大小。分表包括垂直分表和水平分表两种方式。

垂直分表指将数据表按照字段分成多表，每个表存储其中一部分字段。垂直分表减少了每一条记录的数据量，因此在查询时能减少读取的 Block 数和使用的 I/O 数，进一步减少 I/O 争抢的次数，降低锁表的几率。此外，垂直分表简化了表的结构，提高了可维护性。通过将冷门和热门属性使用不同表分离，能提高热门属性操作和查询的速度。但是，垂直分表需要使用冗余的主键并对冗余列进行管理，跨属性查询会引起 Join 操作。此外，垂直分表会让事务变得更加复杂。

水平分表是指将数据表按行的拆分成多表，尽可能使得每个表数据量相当。水平分表能优化单一表数据量过大而产生的性能问题，也能减少 I/O 争抢从而减少锁表的几率。在数据量上，水平分表能简单地达到对大数据量场景的需求，在代码层面应用端只需少量改造。但是水平分表在跨界点需要用复杂的逻辑获取数据，也会遇到性能问题。此外，如果数据持续增长，达到现有分表的瓶颈，需要增加分表，可能会出现数据重新排列的问题。

经过特征提取模块得到的特征数据具有字段较少，行数较多的特点，而且没有大的字段。相比使用垂直分表将不同字段存在不同表中，更合理的策略是采用水平分表，降低单表的行数从而提升插入和查询速度。

水平分表需要根据业务特性找出具体的分割标准。AETA 的原始数据包括信号类型，台站号和时间戳三种确定唯一性的属性，由于电磁扰动和地声探头提取的特征类型有所区别，而且时间戳不一致，第一层按照信号类型进行水平分割。在时间和台站号的选择上，需要结合实际应用场景来判断。由于台站间存在差异性，在数据分析中更关注台站自身特征随时间的变化，往往会先选取一个台站较长时间的各种特征进行异常检测，最后再综合多台站分析地震发生的风险。如果按照月份或季度按时间水平分表，会出现大量跨表查询的场景，所以第二层按照台站进行分割。

特征数据库具体的 E-R 图如图 3.13 所示，其中保存电磁和特征数据的特征表将分别命名为 magn_sta+台站号和 sound_sta+台站号。

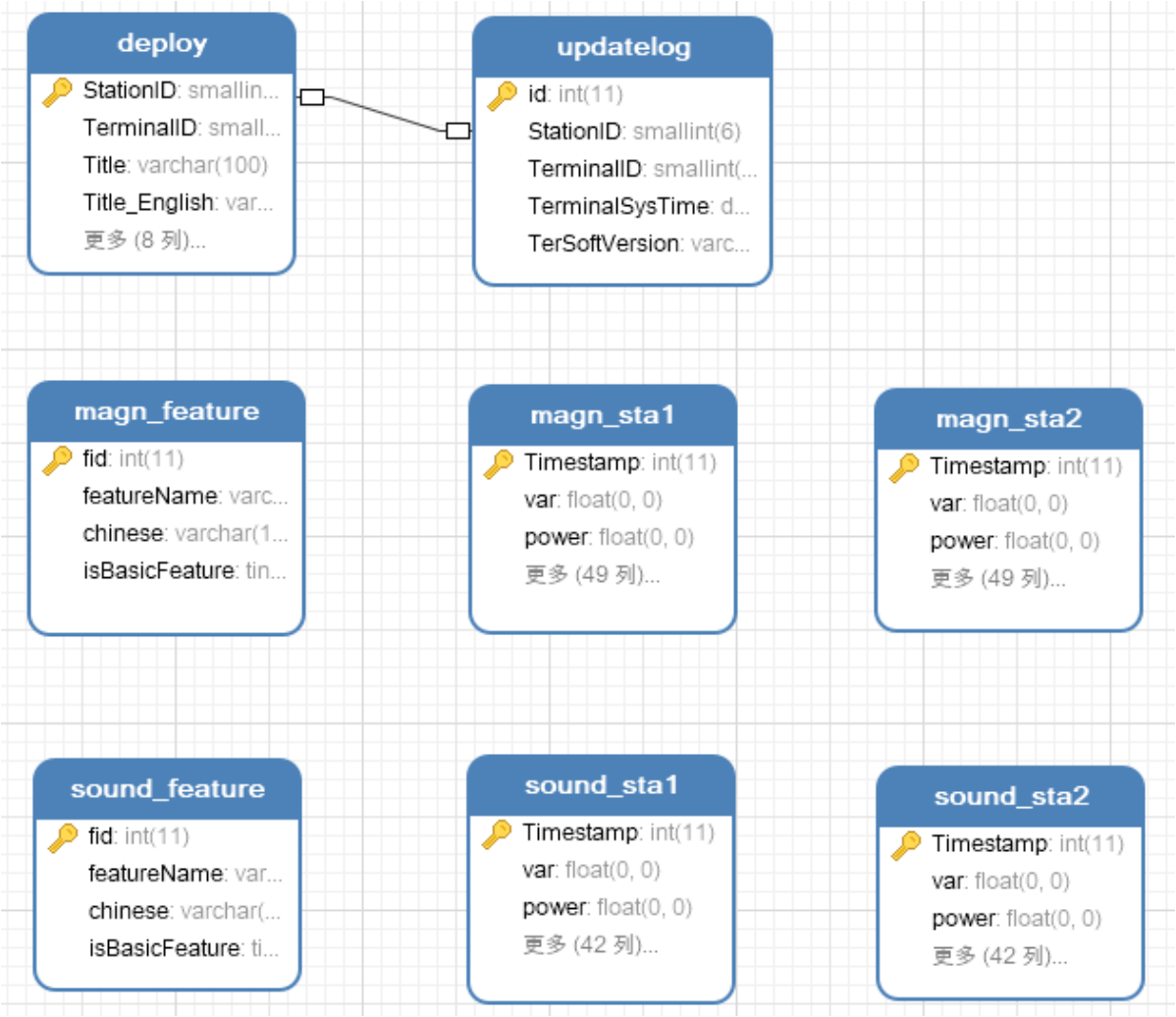


图 3.13　特征数据库 E-R 图

特征数据库中，包含以下 5 个部分。首先 deploy 表记录所有台站相关的信息，以 StationID 为主键，该表将通过每天自动运行的脚本和阿里云数据库的同步。然后是 updatelog 表，记录台站版本更新信息，其中 StationID 与 deploy 表外键约束。接下来是 feature_deploy 表，记录电磁和地声的特征名，最后是水平分表的特征数据表，以时间戳为主键存储。

设计并实现特征数据库后，接下来介绍特征数据库接入模块，对应的框图如图 3.14 所示。

特征数据库接入模块首先从特征数据缓存中获取某一批次预处理数据，从中提取出该批次数据的台站号和信号类型。随着新台站的陆续安装，在特征数据库中可能有新台站没有对应的特征数据表，因此存储数据前需要先查询特征数据库的存在情况。如果不存在，则根据 feature_deploy 表中的特征名信息，创建特征表。接下来从特征数据中提取特征名，并将不在 feature_deploy 表记录的特征视为无效特征，删除对应的数据

并抛出警告。最后将进行数据的录入，考虑到在实际应用中，存在台站补传数据，或监控模块重启某一批次任务等现象。在这种场景下，数据库对应存储的数据需要更新，这种异常需要被考虑和处理。如果已存在对应时间戳的数据，可以选择抛弃数据或者更新数据。如果选择更新，可以采用 MySQL 中的 ON DUPLICATE KEY UPDATE 关键字实现。完成以上流程后，特征数据库接入模块将获取并处理下一批次的数据。

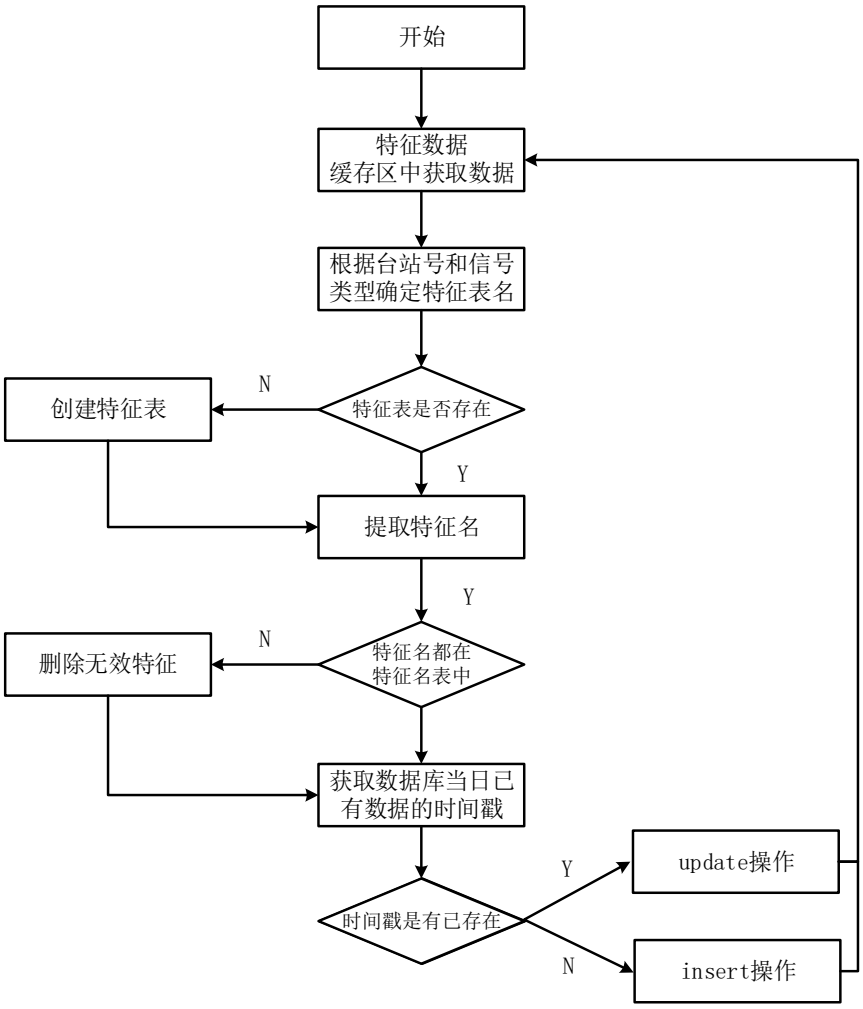


图 3.14　特征数据库接入模块框图

3.6　本章小结

本章针对 AETA 原始数据分析平台中特征生成流部分进行讲述。为了减少各个模块的耦合，同时提高系统的利用率，模块间将使用生产者消费者模型进行连接。

特征生成流首先从原始数据下载器开始，根据任务池中的任务和对应的原始数据文件信息，使用 http 或 PostgreSQL 数据接口从原始数据文件储存系统中下载指定原始数据或从本地磁盘中读取。在预处理模块中，针对原始数据中断电重启异常、数据缺失

异常和脉冲型异常，特定的算法将进行检测和修补。预处理后的数据将在特征提取模块分别提取时域相关的特征，频域相关的特征，以及小波变换相关的特征。其中，电磁扰动信号一共提取 49 个特征，地声信号提取 42 个特征。为了提高特征提取的速度，本章还针对耗时较长的短时能量和短时过零率函数进行优化。最后，特征数据将通过特征数据库接入模块存入特征数据库。由于特征数据行数多，特征数据表按照信号类型以及台站号进行水平分表，提高插入和查询的速度。

第四章　AETA 原始数据异常风险库的设计与实现

上一章介绍了如何将原始数据通过特征生成流生成特征数据并存储在特征数据库中，本章将继续讲述如何针对特征数据进行异常检测，并通过特征选择算法评价异常和地震的相关性，得到异常风险指标，最后持久化存储于异常风险库中。此外，本章还介绍了如何实现对应的数据库接入器，以及展示层如何使用交互式数据可视化框架 Dash，通过波形图，数据分布图，频谱热力图和时空图从多个角度可视化特征数据及其异常。

4.1　异常检测模块的设计与实现

原始数据在时空上具有差异性。对于同一种特征，在不同台站中具有不同的分布，没有直接的可比性。同时，即使针对同一个台站，在较长的时间跨度中稳态也存在区别，更有效的方式是选择同一个台站，从短时跨度的数据出发，关注台站和自身过去状态相比的异常。

异常检测模块主要结合 AETA 数据的特点和日常数据分析的需求，搭建异常检测的框架，并内置一些异常检测算法。搭载的异常检测算法包括单特征的异常检测和多特征的联合异常检测。整个框架具有扩展性，可以通过简单配置添加新的异常检测算法。该模块和后面的其他模块相结合，可以很方便地对特征数据检测异常，并评价异常和地震的相关性。在展示层中还可以可视化异常的时空分布和对应异常与地震的相关性。

4.1.1　异常检测算法概述

异常检测也称偏差(deviation)检测或者离群点(outlier)检测，始于 20 世纪 80 年代，常见的异常检测算法由以下四类。

1、基于统计和概率的方法。基于统计的方法往往需要先对概率的分布进行假设，然后根据给定的假设分布，小概率发生的点将被视为异常。统计分布在有足够先验知识时对单点异常判断较为有效，但对高维数据效果较差。同时假设的分布将直接影响检验效果。不过在基于统计的算法中，滑动四分位不需对统计分布做出假设，直接通过四分位数判断异常。

2、基于距离、密度的方法。该类方法认为相比于正常点，异常点会远离其他点，或者在异常点附近的密度将远远小于正常点。该类方法一般使用硬阈值，缺乏足够的

泛化能力，且时间复杂度一般在$O(N^2)$或以上。

3、基于聚类的方法。根据正常点、异常点和簇的关系，这类方法可以细分为三种。以 DBSCAN 为代表的算法假定正常点不必属于某个特定的集群，但异常点远离所有簇的簇心，以 SOM, K-means 为代表的算法假定正常点靠近距离中心而异常点远离中心，以及 find_CBLOF 为代表的算法假定正常点的集群规模和密度远大于异常点的集群。

4、基于集成学习的算法。这种方法以周志华提出的孤立森林[53]为代表。孤立森林假设总体样本中只要少数异常点且与正常样本有较大差异性。这些异常样本在随机划分的过程中很快到底叶子节点，树深较低。通过多次随机划分并取平均，离密度高而稀疏的点被判定为异常点。可并行生成且线性复杂度的特点让孤立森林可快速实现异常判断。

除了通用的异常检测方法，针对时间序列的特点，还有以下四类用于时间序列的异常检测方法。

1、基于时间序列预测的方法。这类方法将真实值与拟合后的值误差较大的点视为异常点。预测方式包括回归拟合的算法，如 AR，MA，将两者结合的 ARMA 以及先进行差分的 ARIMA[54]。此外，还有 Holt-winter 算法以及利用深度学习中 RNN，LSTM[55]等包含时间顺序关系的网络。

2、基于时间序列分解的异常检测算法。这类算法将时间序列分解成多个分量，在这些分量中，如果表示残差或随机误差的分量大于阈值，对应的点将被视为异常。分解时间序列往往需要较大运算量，常用的算法如 PCA。

3、基于数据频率的编码方法。这类算法根据时间关系的序列出现的频率来判断异常，最经典的包括马尔科夫模型（MM）以及隐马尔科夫模型（HMM）[56]。针对 HMM 需要大量样本的问题，可以使用半监督学习的方法或增量学习的方法来改进[57,58]。针对时间复杂度过高的问题，可以提出低频特征序列[59]。

4、基于免疫的算法。这类算法基于生物免疫识别中的自我-非我识别原理，让自我空间的检验器集尽量被正样本覆盖，然后使用检测器判别系统的运行状态。在模型学习时这类方法需要大量的正样本，适用于异常样本未知的情况。具体的算法有负选择算法(NSA)[60]。在其基础上可使用模拟退火算法，欧氏距离，粒子群算法，网格的特征空间等方式解决检测器数量和对非我空间覆盖之间的矛盾[61-64]。

异常检测模块从特征数据库获取特征数据，并进一步地搭建了一个异常检测框架。在这个框架中，集成了项目组中已取得较好效果的低复杂度算法，具体为基于单特征统计的滑动四分位法和 ksigma 算法，以及基于多特征集成学习的孤立森林算法。框架具有扩展性，除了已有的方法，对于未来新发现有效的算法，可以很方便地在框架中添加并用于检测。

4.1.2　异常检测模块算法与框架

异常检测模块的框图如图 4.1 所示。

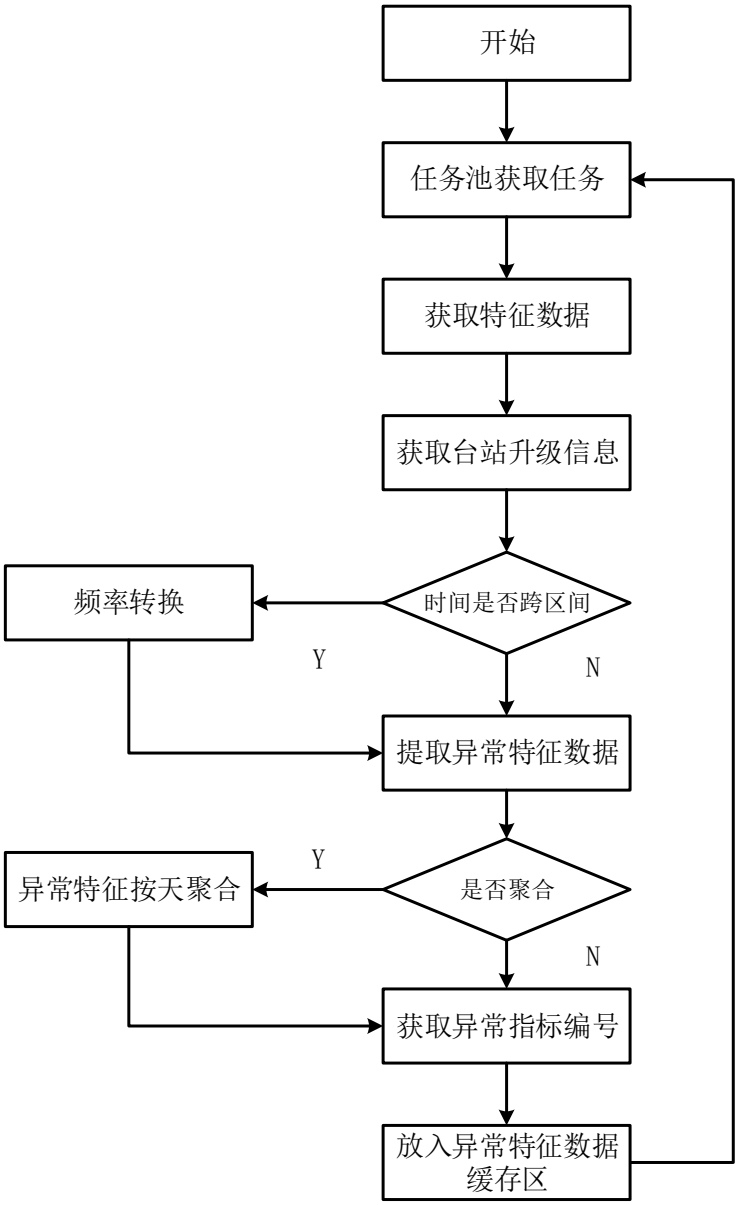


图 4.1　异常检测模块框图

从任务池获取到某一项任务之后，异常检测模块将通过调用数据库接入器的接口，获取对应的特征数据。由于在 AETA 系统的版本升级变更中，数据采集的时间抽样间隔有 3 个不同的值，需要通过频率转换和重新采样的方法统一采样率，下一步将从数据库中获取台站历史升级的信息，以方便判断获取的特征数据是否存在跨采样区间的情况。如果存在，将采样频率转换到同一个值。

转换方法包括向上采样和向下采样，前者通过增加数据点，从低采样率变成高采样

率，常用插值的方式来实现。后者通过聚合的方式降低采样率，常常采用箱体抽样的方式。在具体实现上编写了一个工具类，向上采样的方法包括线性插值法，三次样条插值法和 sinc 函数内插。向下采样主要通过各种聚合函数，包括均值，最小值，最大值。具体的调用通过传入的参数来选择。

完成频率转换之后，根据传入的异常检测算法，调用对应的函数得到异常值并根据传入的设置参数决定是否对异常进行聚合操作。下一步对于每种异常检测算法，使用数据库接口获取对应的异常编号，最后将异常编号和异常值一并放入异常风险数据缓冲区中。

4.2 异常评价模块的设计与实现

4.2.1 特征评价算法

在实际的数据分析场景中，往往采用二分类来进行预测，即有震和无震。要评价某个异常风险指标和地震的关系，首先需要根据地震列表进行打标得到地震标签。在孕震过程中，异常会在震源区及其附近不同程度地出现，这些异常往往出现在地震前后，所以除了需要将地震当天的标签置 1 以外，还需要根据地震的震级，台站的震中距，给不同台站发震前后不同时间范围的标签记为 1。具体打标方式需要先在数据库记录，然后程序根据标签方法 ID 从数据库读取对应的配置并进行打标。表 4.1 显示了默认的打标参数，其中里面的天数表示和发震时刻相对时间，例如-3~0 表示将震前 3 天到地震当天打标为 1。

表 4.1 默认打标参数

台站震中距 (KM)	有感地震（MS3~4.5）	中强震（MS4.5~6）	强震（MS6 以上）
＜100	-3~0	-5~0	-7~0
100~300	-2~0	-3~0	-5~0
300~500	无	-2~0	-3~0

打标函数的函数名和打标方法将存在数据库的 config_label 表中。

除了默认的打标函数外，还可以自定义打标函数，同时将函数名和注释录入数据库。除了二分类的打标方式外，也可以设置连续值标签，需要在 config_label 中江 isDiscrete 设为 0，标注为连续值标签。

对于提取的异常风险指标，异常评价模块集成了 3 个方面的算法进行评价。分别为方差分析，AUC 指标和时序叠加分析。

1、方差分析

对于标签为离散值的分类任务，评价异常指标和标签的相关性可以使用方差分析的方法。根据地震标签，可以将异常指标分为S^+和S^-两个集合。其中S^+为标签为 1 异常指标的集合，S^-为标签为 0 的集合。若S^+和S^-差异越大，表示该异常指标越能有效区分有无地震。假设$H_0: \mu_{S^+} = \mu_{S^-}$，则检验统计量$f$为：

$$f = \frac{S_A/(r-1)}{S_B/(n-r)} \tag{4.1}$$

其中S_A, S_E分别是组间和组内方差，n 为样本数，r为类别数。根据f能得到显著性p值，p 值越小，代表正负集合差异越大，异常指标和标签的相关性越高。

2、AUC 指标

在评价二分类的准确性的指标中，AUC (Area Under Curve)是个重要的参数，由 ROC (Receiver Operating Characteristic) 曲线得到。ROC 曲线最早应用于雷达信号检测领域，通过连结不同阈值时的假正率 (FPR) 和真正率 (TPR) 而形成。AUC 指的是 ROC 曲线与坐标轴围成的右下方面积，它本质表示随机抽取一个正样本和一个负样本，正样本得分高于负样本得分的概率，即使样本正负标签不平衡，也能有效地做出评价。

3、时序叠加分析

在天体、地震研究等领域，广泛应用时序叠加分析 (SEA) 来从统计分析方的角度验证某特征是否与特定事件有显著关系[65-69]。时序叠加分析通过选取 n 次地震发震$\pm t$天的数据的异常值进行叠加得到地震相关值序列EV_i，然后随机抽要得到随机从所有天数随机选取 n 天用同样的方式叠加$\pm t$天的数据得到背景序列BG_i。重复进行 M 次抽样，得到BG_i的均值μ_i和方差σ_i。如果EV_i存在某一天的值大于$\mu_i + 2\sigma_{\mathrm{i}}$，则认为该特征和地震有显著相关性。为了量化衡量两者的相关性，本文采用EV_i在$\mu_i + 2\sigma_{\mathrm{i}}$曲线上的面积作为 SEA 得分。

4.2.2　异常评价模块框架

异常检测模块的框图如图 4.2 所示。

异常检测模块首先根据打标参数，调用数据库接口，获取对应的打标编号。如果没有，根据需要新建一个编号或抛出异常。在每一批次的数据中，异常检测模块将从特征数据库中获取一个台站指定时间的数据。下一步将获取台站周边对应时间段的地震数据，对饮时间段为特征数据开始时间早 15 天到特征数据结束时间晚 15 天，下一步根据打标参数，结合发震时间生成对应的地震标签。得到异常风险指标和地震标签后，接下来通过方差分析，AUC 指标和时间叠加分析三个角度评价异常风险指标和地震的相关性，汇总后放入异常风险评价指标缓冲区中。缓冲区中的数据将通过其他线程将异常风险指标入库。

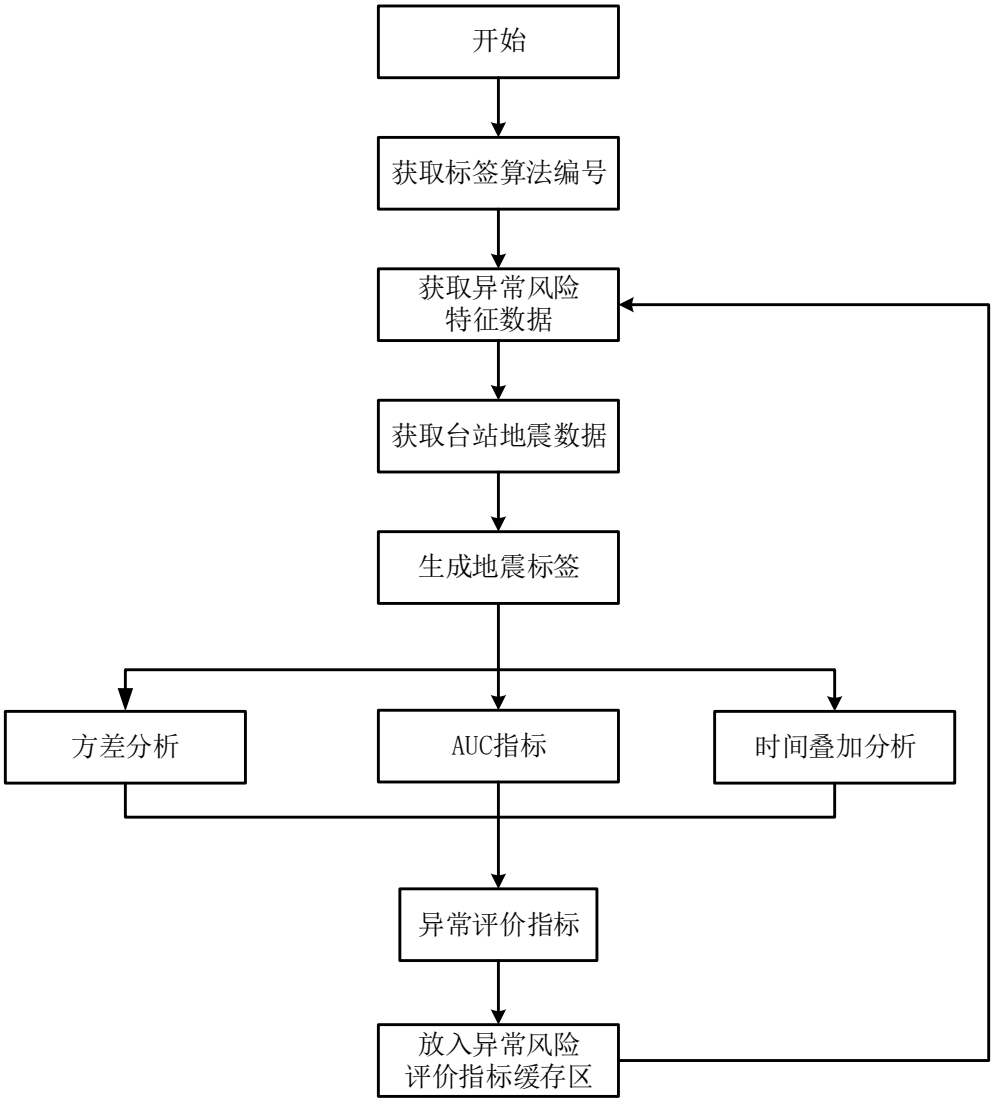


图 4.2　异常评价模块框图

4.3　异常风险库的设计与实现

异常风险库主要用于存储台站信号检测到的异常，以及每一种异常和地震的相关性。因此，异常风险库围绕标注异常类型的 abnormal_type 表来进行设计，并通过 abnormal_log 和 abnormal_score 来记录出现的异常以及对应和地震的相关性得分。此外，由于存在多种异常检测算法和地震打标的模式，需要分别设计表格来记录方法和具体的配置。为了提高查询速度和减少存储的数据量，异常日志将存储异常量按天聚合值，同时对于无数据或数据太少的天数不做存储。对于后者，为了和无异常做区分，将用一个专门的表存储台站数据太少的天数。此外，电磁和地声通过水平分表分开。各数据表 E-R 如图 4.3 所示，其中钥匙形状对应的字段表示表的主键，相连的线表示外键约束。

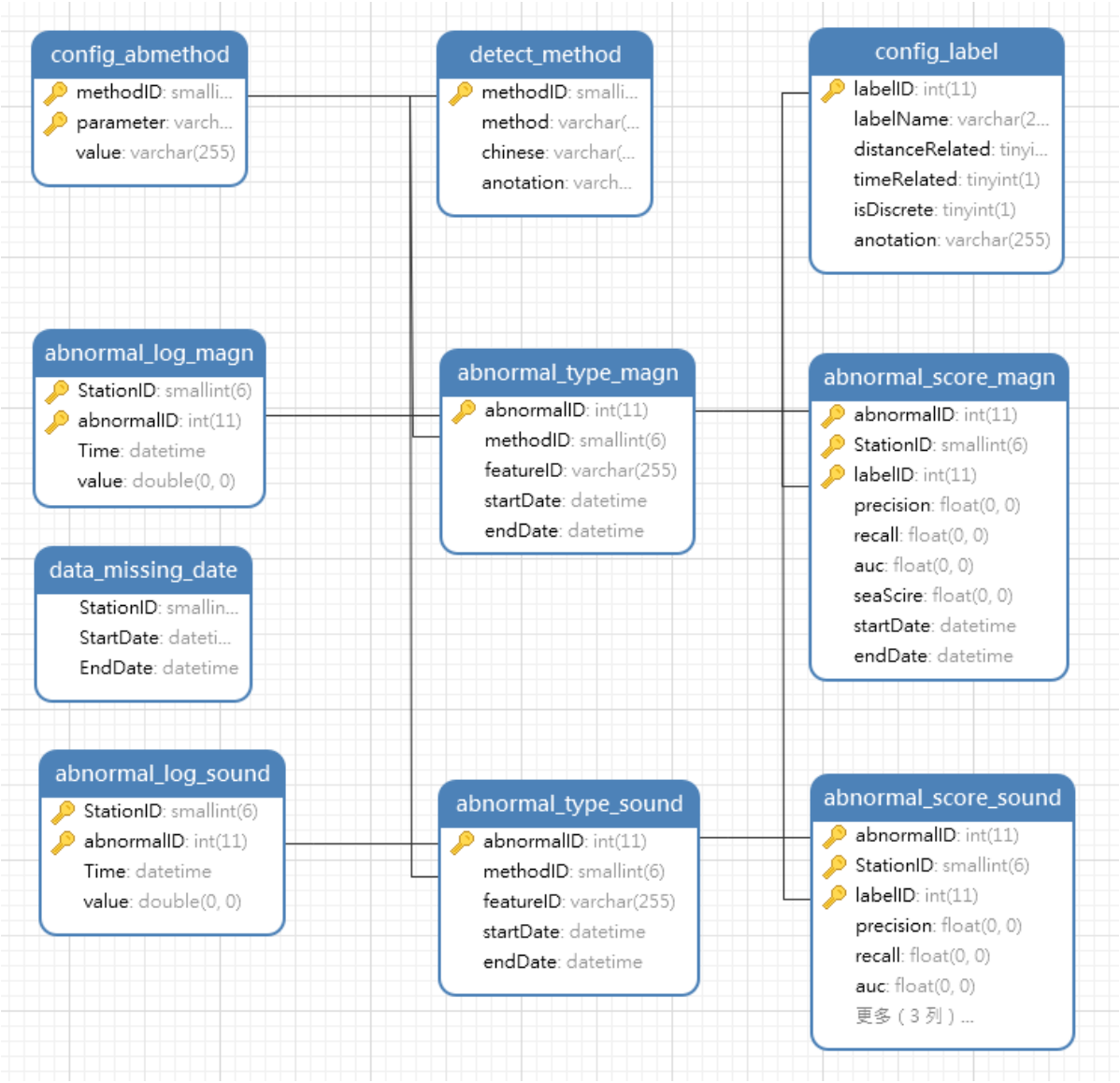


图 4.3　异常风险库 E-R 表

4.4　数据库接入器的设计与实现

数据库接入器主要实现持久化存储的 MySQL 数据库和数据处理层的交互，避免数据处理层对底层数据的直接操作，提高数据安全性。根据对特征数据流和异常风险库各个模块的设计，特征数据库接入器有以下 8 点需求

1、从阿里云的数据库中获取特定台站的重启信息，用于数据预处理模块判断断电重启异常。

2、判断某个台站对应的特征表是否存在，如果不存在，按照特征名表中的基本特征新建特征表。

3、 从特征名表中获取电磁扰动或地声数据有效的基本特征，从而删除特征提取模

块错误运行或其他原因提取的非基本特征。同时将数据分析中发现的有效的衍生特征加入特征名表中。

4、将特征数据存储到特征表中。没有主键冲突时，采用插入模式，否则可根据选择采用更新模式或忽略。

5、从特征数据库读取特征数据，获取台站升级信息，用于对不同时间粒度的数据进行频率转换。

6、获取异常指标，异常检测算法和打标算法对应的 ID 号。如果不存在，则在对应的表中新建 ID。

7、 将异常风险指标插入异常日志中，并能根据台站号，信号类型，时间范围以及异常指标 ID，获取或写入对应的异常指标。

8、将异常风险评价指标插入异常得分表中，并能从异常得分表中获取异常风险评价指标数据，以用于展示层可视化。

根据以上 8 点需求，数据库接入器一共有 21 个接口，每个接口的名称，参数和描述具体如表 4.3 所示。

表 4.3　数据库接入器接口

序号	接口名称	接口描述
1	create_coon(conf)	创建数据库连接
2	del_coon(conf)	断开数据库连接
3	get_basic_feature(signalType)	获取基本特征
4	add_feature_table(signalType, featureDict)	插入非基本特征
5	create_feature_table(stationID, signalType)	创建特征表
6	get_update_log(stationID, timeRange)	获取台站升级信息
7	get_restart_log(stationID, timeRange)	获取台站重启信息
8	get_station_feature_data(stationID, signalType, timeRange, featureList)	从特征表中获取台站的特征数据
9	put_station_feature_data(data_df, stationID, signalType, mode="insert")	往特征表插入/更新台站的特征数据
10	add_abnormal(feature, method, signalType)	添加异常指标
11	get_abnormalID(feature, method, signalType)	获取异常指标 ID
12	add_abnormal_method(name, parameterDict)	添加异常检测方法

续表 4.3　数据库接入器接口

13	get_abnormal_methodID(name, parameterDict, add=False)	获取异常检测方法 ID
14	add_lable_method(name, parameterDict)	添加打标算法
15	get_lable_method(name, parameterDict)	获取打标算法 ID
16	add_missing_date(stationID, datestr, signalType)	添加台站数据缺失情况
17	get_missing_date(stationID, timeRange, signalType)	查询数据缺失情况
18	put_abnormal_score(data_df)	插入异常指标和地震相关性
19	get_abnormal_score(abnormalID, stationID, lableID, timeRange)	获取异常指标和地震相关性
20	put_abnormal_score(signalType, stationID, labelID, timeRange, score_dict)	插入异常数据
21	get_abnormal_log(stationID, signalType, timeRange, threadDict={})	获取异常数据

4.5　展示层的设计与实现

展示层用于将数据处理层的数据以可视化的方式展示出来，包括台站的特征数据，异常风险指标，异常风险对应的原始数据波形，以及异常风险指标的评分。展示层可以通过读取数据处理放在缓冲区的数据调用接口生成为 html 文件并查看，也可以通过读取持久化存储在存储层的数据并通过网页查看并进行交互。具体可视化方法包括波形图，数据分布图，频谱热力图以及时空图。

波形图反映某变量随着时间变化的过程，主要用于展示各个特征数据和某个异常对应原始数据随时间的变化。波形图需要支持不同的时间跨度，如日尺度，周尺度月尺度等。当选取数据点过多时，可以根据参数设置是否向下采样进行聚合操作。同时，针对地震前兆分析这个实际需求，波形图可以根据需要显示地震事件，以更直观地展示波形和地震之间的关联。此外，波形图还需要支持同时显示多条数据，以方便对比。如同一特征当日数据和昨天，一周前同一时间段的对比，同一台站同一时间段不同特征的对比，以及同一时间段不同台站同一特征的对比。

数据分布图主要显示特征的分布，包括直方图，箱型图和密度图。根据需要，可以添加数据的一些统计量，包括均值，方差，偏度和峰度。数据分布图除了显示一维数据的分布外，也可以传入两个特征，显示它们的联合密度分布，以从特征交叉的角度给人

直观展示，方便挖掘更多的信息。

频谱热力图使用热力图的方式可视化原始数据在不同频段的能量的特征，展示随着时间的推移，信号频谱的变化。在地震的孕育过程中，往往在某系特定的频段发出能量，可以通过频谱热力图发现这些异常。除了显示各频段能量的绝对值大小外，也可以选择显示各频段能量占所有能量的百分比，展示相对值的变化情况。

时空图用于展示异常在空间上的分布，以及显示各地区的异常随着时间变化出现和消失。时空图以天为粒度，展示当前时段不同台站特征数据中检测到的异常，并通过不同的符号或颜色展示异常的种类或异常风险与地震的相关性。当鼠标悬浮在异常点时，可以显示该异常和地震的相关性。时空图可以联合多台站进行分析，克服单台站的容易误报的缺点，让分析者对总体的异常情况有更清晰的了解，更好地进行地震的预测。

实现方面使用 MIT 协议开源的 Dash 进行开发。Dash 是一个用于构建 Web 应用程序的高效 Python 可视化框架，其基于 Flask，Plotly.js 和 React.js，用户能够使用 Dash 构建高度自定义界面的数据可视化应用程序，从而进行数据分析和数据探索。网页页面由波形图，数据分布图，频谱热力图和时空图四个部分组成，每个部分首先选择台站，信号分量，时间等参数，点击查询生成图像。图像可以查看缩放筛选保存，当鼠标悬停数据点时可以展示信息。此外，由于 Dash 里面的图像均通过 plotly 构建，画图的接口单独封装成一个类，除了直接通过网页查看外，其他程序也能调用接口生成 html 文件保存到本地，从而方便分析者日常查看和根据需要自行调用。

4.6 本章小结

本章主要针对 AETA 原始数据分析平台中异常风险库部分进行讲述。由于台站数据存在时空上的区别，特征数据需要进一步地通过异常检测，提取其中的异常来进行地震预测。本章首先整理了异常检测算法方面常用算法，并在异常检测模块集成了项目组中应用已有较好效果的低复杂度算法，具体为基于统计的单一特征异常检测算法滑动四分位法和 ksigma，以及基于多特征集成学习的孤立森林算法。检测出的异常风险指标将通过方差分析，AUC 指标和时间叠加分析三个方面评价异常和地震的相关性。本章还设计了持久化储存的异常风险库，针对数据分析平台各个环节对数据库的需求，详细介绍数据库接入器 21 个接口的含义的使用方法。最后，特征数据以及其异常和分布将通过展示层以波形图，数据分布图，频谱热力图和时空图多个角度可视化。

第五章　AETA 原始数据分析平台的测试与展示

之前的章节分析了 AETA 原始数据分析平台的需求，并自上而下地设计与实现了原始数据分析平台，本章将进行对应的分析测试与展示。特征生成流的数据量大，且提取的特征会直接影响后面数据分析，除了对正确性，功能性测试外，还将重点关注时间开销。另一方面，异常风险库相对数据量较小，更多关注其功能和最后的可视化层。

5.1　测试环境

根据日常实际使用的环境不同，本章测试使用的设备包括实验室的台式机以及实验室的服务器，两种设备具体的配置如表 5.1 所示。

表 5.1　设备环境

设备	服务器	台式机
CPU	E5-2640	i3-3240
主频	2.40GHz	3.40GHz
内存	128G	32G
操作系统	Centos 6.9	Win10 1903 版本
python 版本	3.7.7	3.7.7

其中绝大部分测试在服务器上进行。由于实际日常的数据分析的使用场景下，一般从台式机获取硬盘中存储的数据，因此针对原始数据从本地读取的测试在实验室的台式机中进行。

5.2　AETA 原始数据特征生成流

5.2.1　原始数据下载器

原始数据下载器是特征生成流的起始环节，下载的准确性和速度将直接影响提取特征的有效性和整体特征生成流的运行时间。本小节将分别从下载数据的准确性和下载速度两方面对原始数据下载器进行测试。

验证数据下载的正确性最直接有效的方法是比较下载的数据是否与原文件具有相

同的哈希值。由于原始数据文件存储系统并没有提供每条数据的哈希值，这里采用间接的验证方式。对于每条原始数据，阿里云服务器将提取特征值并存放到云端数据库中。如果下载的数据正确，采用相同的算法对提取特征将会得到完全一致的特征值。本文将使用现有均值的算法提取特征并和云端数据库存储的特征进行对比，验证下载数据的正确性。

选择 90 号茂县测点 2020 年 3 月 1 日~2020 年 3 月 31 日的数据进行对比，结果如图 5.1 所示。红线表示直接计算浮点数格式的结果和云端数据库的误差。由于数据库以 DEMICAL 形式存储，并只保留 4 个小数，这导致了存储存在精度损失。对结果取 4 位小数后，误差如绿线所示。可以看到提取的特征和数据库一致，表明原始数据下载器能准确地实现下载功能。

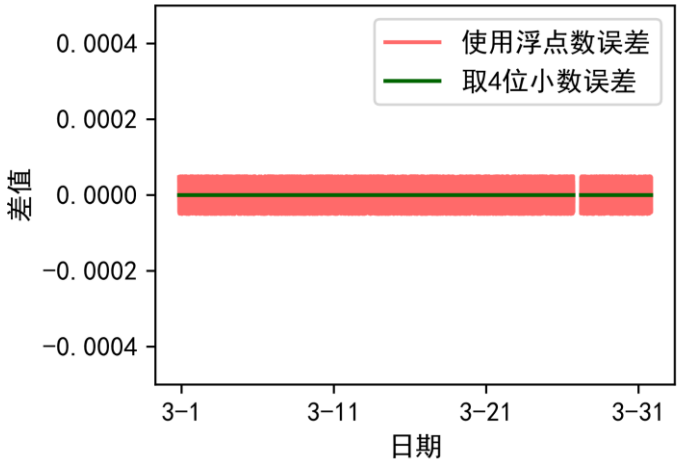


图 5.1　对下载的数据提取现有均值特征与数据库存储值的对比

原始数据文件存储系统会将高频访问的数据放入缓存，以提高热门数据的查询速度，并减少对硬盘阵列的查询次数。这导致对于同一批次的数据，相比第一次的查询，后面的查询速度会明显更快。为了使结果具有可比较性，对从 html 接口获取，从 PostgreSQL 接口获取和从本地已存储文件读取这 3 种不同的方式，将选用同一台站 3 段相同时间跨度的时段，分别用这 3 种方式单线程下载。选择的台站为 38 号冕宁防震减灾局，具体的时段和结果如表 5.2 所示。

表 5.2　原始数据下载器运行时间

方式	时段	均值 (s)	最大值(s)	最小值(s)	标准差(s)
http	2020.03.01~2020.03.31	21.2096	30.4643	9.2952	78.7359
PostgreSQL	2020.01.01~2020.01.31	20.7984	26.3157	15.0627	3.9582
本地读取	2019.12.01~2019.12.31	1.1712	1.5570	0.9034	0.0290

可以看到，从 PostgreSQL 接口和 http 接口下载下载单台站单特征单分量的数据均在 21s 左右。由于在查询数据时，对于新的数据，主服务器需要与从服务器通信，这两种下载方式都取决于服务器间的通讯速度和对外通讯网速。从本地获取的速度平均只需要 1.17s，远快于下载的速度。该速度取决于磁盘读取的速度与 CPU 处理数据的速度的上限。原始数据下载器能满足下载原始数据的需求，同时在本地存在数据时能加快获取数据的速度，减少对服务器的访问。

5.2.2　数据预处理模块

本小节主要展示数据预处理模块对断电重启异常、数据缺失异常和脉冲型异常的检测和处理效果，并测试数据预处理模块整体的运行时间。

图 5.2 展示了 38 号冕宁防震减灾局 2019 年 11 月 1 日~2019 年 11 月 31 日现有均值特征的波形，其中出现了 7 次明显的高值异常。图 5.3 和图 5.4 分别展示了这段时间终端的重启记录和 11 月 27 日采集到的现有均值特征。从图中可以看到，数据预处理模块能有效地检测出前 6 次和重启相关的异常，而 11 月 27 日的第七次异常则为探头运行中采集到的异常信号。

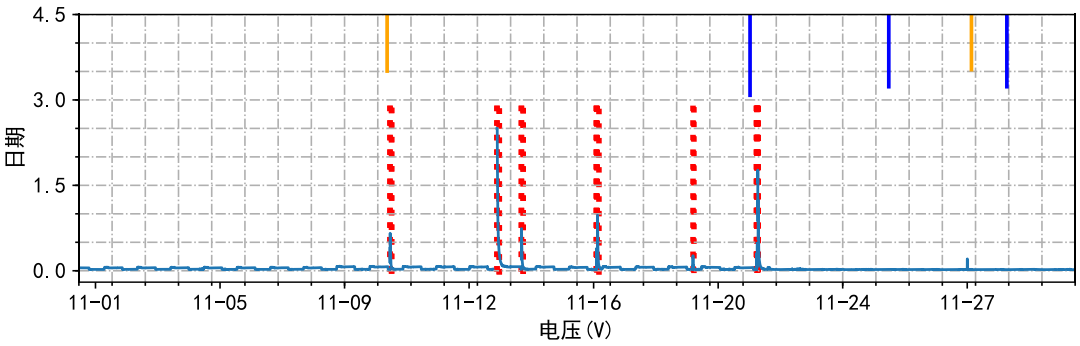


图 5.2　断电重启异常识别效果

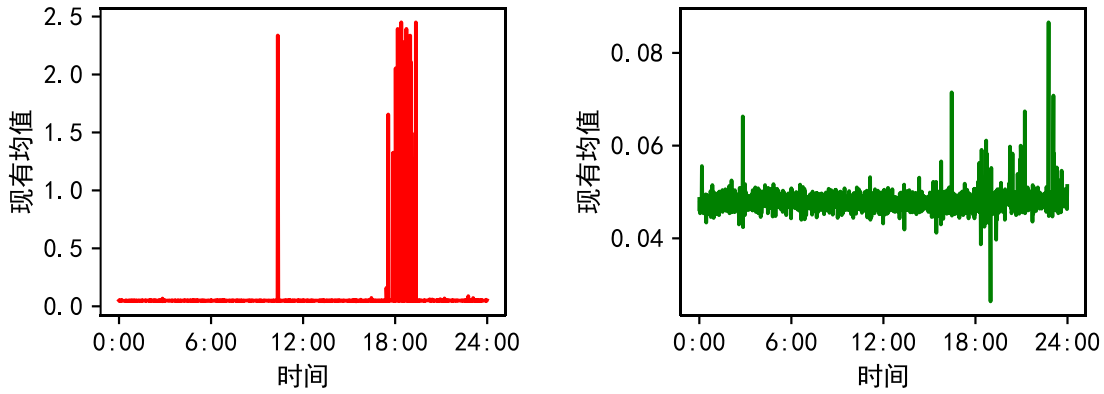
id	TerminalID	ServerSysTime	TerminalSysTime	TerStartTime	TerSoftVersion	SDStatus
66753131	38	2019-11-10 08:52:21	2019-11-10 08:52:20	2019-11-10 08:49:17	3.3.2	0
66986510	38	2019-11-13 14:10:11	2019-11-13 14:10:08	2019-11-13 14:07:06	3.3.2	0
67040398	38	2019-11-14 07:49:24	2019-11-14 07:49:22	2019-11-14 07:46:19	3.3.2	0
67203238	38	2019-11-16 13:59:23	2019-11-16 13:59:21	2019-11-16 13:56:18	3.3.2	0
67204236	38	2019-11-16 14:19:51	2019-11-16 14:19:42	2019-11-16 14:16:39	3.3.2	0
67205345	38	2019-11-16 14:42:30	2019-11-16 14:42:27	2019-11-16 14:39:25	3.3.2	0
67408480	38	2019-11-19 11:50:17	2019-11-19 11:50:15	2019-11-19 11:47:11	3.3.2	0
67549591	38	2019-11-21 09:43:20	2019-11-21 09:43:18	2019-11-21 09:40:16	3.3.2	0
67551097	38	2019-11-21 10:12:38	2019-11-21 10:12:35	2019-11-21 10:09:32	3.3.2	0
67552323	38	2019-11-21 10:36:59	2019-11-21 10:36:57	2019-11-21 10:33:54	3.3.2	-1
67553041	38	2019-11-21 10:51:18	2019-11-21 10:51:16	2019-11-21 10:48:13	3.3.2	0

图 5.3　冕宁防震减灾局 2019 年 11 月重启记录

Time	TerminalID	ProbeID	average
1574848816	38	30021	0.0201
1574848876	38	30021	0.0277
1574848937	38	30021	0.0452
1574848997	38	30021	0.0907
1574849057	38	30021	0.2084
1574849117	38	30021	0.0211
1574849177	38	30021	0.0211

图 5.4　2019 年 11 月 27 日部分冕宁防震减灾局现有均值特征

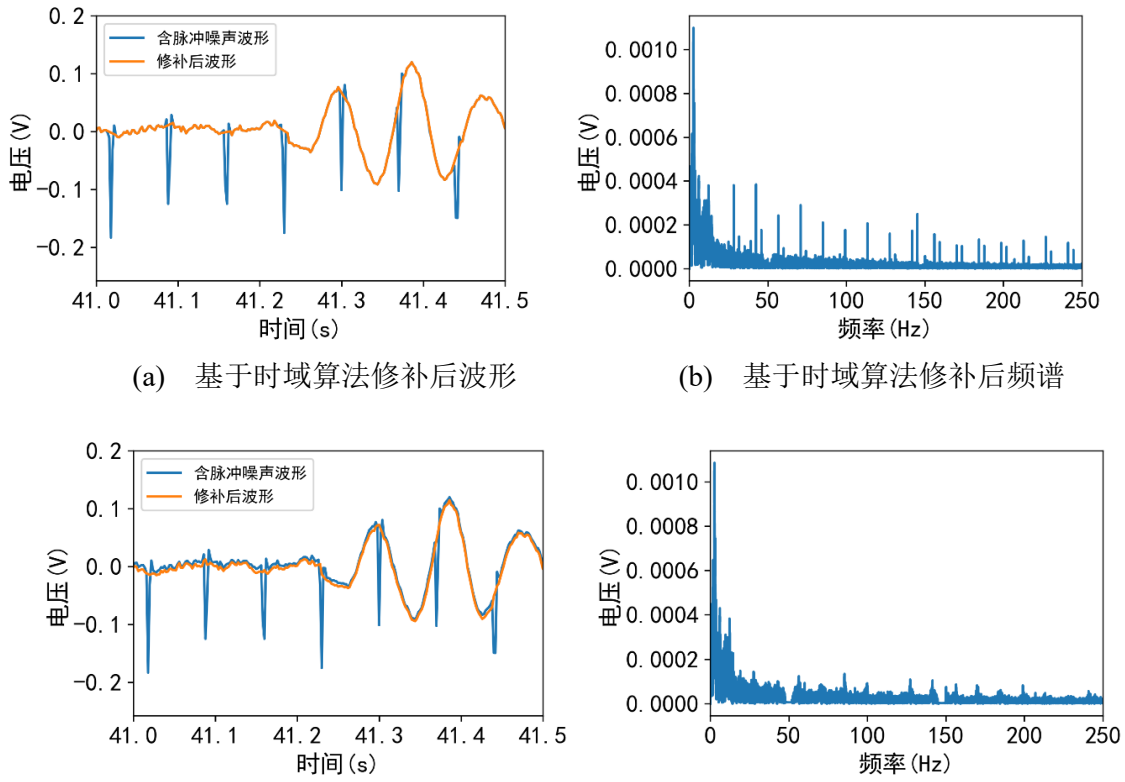
对于数据缺失导致的异常，图 5.5 展示了乐山防震减灾局 5 月 20 日的地声现有均值数据处理前和处理后效果的对比。在处理之前，当天一共出现 28 次数据丢失异常，其中在 17:26 分到 19:21 的时间段甚至产生了数据缺失导致的连续异常区间。处理后得到图(b)的结果，可以看到，经过数据预处理模块的处理，这 28 次异常均被准确地识别并去除。



(a)　处理前的现有均值　　(b)　处理后的现有均值

图 5.5　数据缺失异常修复效果

最后是对脉冲异常的检测和修复。对于含有脉冲异常的信号，图 5.6 分别展示了基于时域和基于频域两种算法修复后的波形和频谱。从图中可以看到，两种算法均能检测出脉冲异常并进行对应的修复，但两者的效果略有区别。基于时域的算法对检测的脉冲异常波形使用插值处理，在时域中能很好地保存原来的波形形状，但在频谱未能完全去除脉冲导致的异常，依然存在等间隔的峰值。相比而言，基于频域的算法能很好地抹除脉冲异常对频域中的影响，但在反傅里叶变换得到的时域在非脉冲异常区间和原来的波形略有区别。运行时可以根据预设的参数选择所需的算法，默认情况下使用时域修补算法并将预估的频率传到后面的模块。



(a)　基于时域算法修补后波形　　(b)　基于时域算法修补后频谱

(c)　基于频域算法修补后波形　　(d)　基于频域算法修补后频谱

图 5.6　脉冲异常修复结果

对特征提取模块整体运行时间的测试结果如表 5.3 所示。由于脉冲异常的分布在台站中具有集中性，而且耗时较长，为了更有代表性地测试运行时间，本文选取了两种极端情况。具体的选取为日常出现脉冲异常的 151 号鲁甸地震台，以及从未出现脉冲异常的 90 号茂县测点，数据范围为 2019 年 12 月 1 日~2019 年 12 月 31 日的地声原始数据。此外，断电异常只需要根据现有特征数据整体运行一次即可。从表中可以看到，数据预处理模块时间能在 6.3~8.8 秒的时间内处理完数据，小于没有本地数据时原始数据下载器的运行时间，满足实际需求。

表 5.3　特征提取模块运行时间

台站号	断电异常 (月/s)	数据缺失异常 (s)	带通滤波 (s)	脉冲异常 (s)	其他时间 (s)	除断电异常总时间(s)
151	0.5072	0.0867	2.9550	5.7776	0.0228	8.8420
90	0.5249	0.0862	2.8648	3.4012	0.0026	6.3548

5.2.3 特征提取模块

特征提取模块对预处理后的数据提取特征，是耗费最多计算资源的模块。此外，在3.4.4 小节中介绍了如何加速耗时较长的短时相关特征，本小节将重点针对运算时间进行测试。

选择 90 号茂县测点 2020 年 3 月 1 日地声原始数据进行测试。当天地声原始数据一共 1440 条，总共大小为 84.375MB。当窗口长度为 400 个点时，选择不同的滑窗长度，每种长度重复测试 10 次并取平均，不同算法提取短时能量的运行时间如表 5.4 所示。

表 5.4　不同算法提取短时能量特征的运行时间比较

滑窗长度	遍历法 (s)	矩阵法 (s)	优化算法 (s)	与遍历法对比	与矩阵法对比
200	4.9033	0.8184	0.5949	-87.87%	-27.31%
100	9.3098	2.8366	1.0927	-88.26%	-61.48%
80	12.5070	3.7549	1.3458	-89.24%	-64.16%
50	19.5032	6.7813	2.1422	-89.02%	-68.41%

从表中可以看到，优化后的算法比遍历法总体上减少 87.87%~89.24%的运行时间，比矩阵法减少 27.31%~68.41%的运行时间。相比于遍历运算，矩阵法和优化后的算法均能大幅减少时耗。伴随着滑窗长度的减少，总体帧数的增加，遍历运算和优化后的算法时耗呈线性增长，而矩阵法的时耗快于线性增长值，这主要由于矩阵法需要申请更多的内存。优化后的算法能大幅提升短时能量特征的提取速度并比矩阵法更高效。

双阈值短时过零率使用和短时能量相同的数据进行测试，结果如表 5.5 所示。实验结果表明，使用遍历法计算双阈值短时过零率平均需要 215.8s，对 numpy 遍历操作速度过慢。采用矢量运算后时间减少 99.82%至 0.387s，满足实际需求。

表 5.5　不同算法提取双门限短时过零率特征的运行时间比较

算法	平均值(s)	最大值(s)	最小值(s)	标准差(s)
遍历法	215.8195	216.1565	215.2486	0.3218
矢量计算	0.3867	0.3874	0.3861	0.0005

最后对特征提取模块整体进行测试。选择 90 号茂县测点 2020 年 3 月 1 日~2020 年 3 月 31 日的地声数据进行特征提取，测试每日平均运行时间。统计过程中去除数据部

分缺失的 3 月 26~27 日，即求 29 天的均值，结果如表 5.6 所示。

表 5.6　特征提取模块运行时间比较

算法	时域相关(s)	频域相关(s)	小波相关(s)	其他环节(s)	总时间(s)
遍历法	234.8134	2.1828	50.0950	0.0122	287.1034
优化短时过零率	15.4389	2.1602	51.5785	0.0112	69.1889
矩阵法	5.6252	2.1800	19.6333	0.0111	27.4496
优化算法	5.2727	2.1974	10.3506	0.0110	17.8317

结果表明，相比于都采用遍历法，优化短时过零率算法能减少 75.9%的耗时，而进一步优化每条数据调用 5 次的短时能量函数后，耗时能在只优化短时过零率的基础上减少 74.2%至每天 17.83s。运行时间的减少主要集中在时域相关的特征和小波相关的特征，相比而言频域相关的特征以及其他环节基本不变。这是由于在特征提取时，时域相关的特征需要调用 1 次短时能量函数，1 次短时过零率函数，小波相关特征需要调用 5 次短时能量函数。通过对相关算法的优化，特征提取模块能实现对特征的高效提取。

5.2.4　数据库接入器

选择 90 号茂县测点 2017 年 6 月 11 日~2020 年 4 月 30 日的数据提取特征并使用数据库接入录入数据库，结果如图 5.7 所示，可以看到，特征数据被能有效存入数据库，从而实现持久化存储。

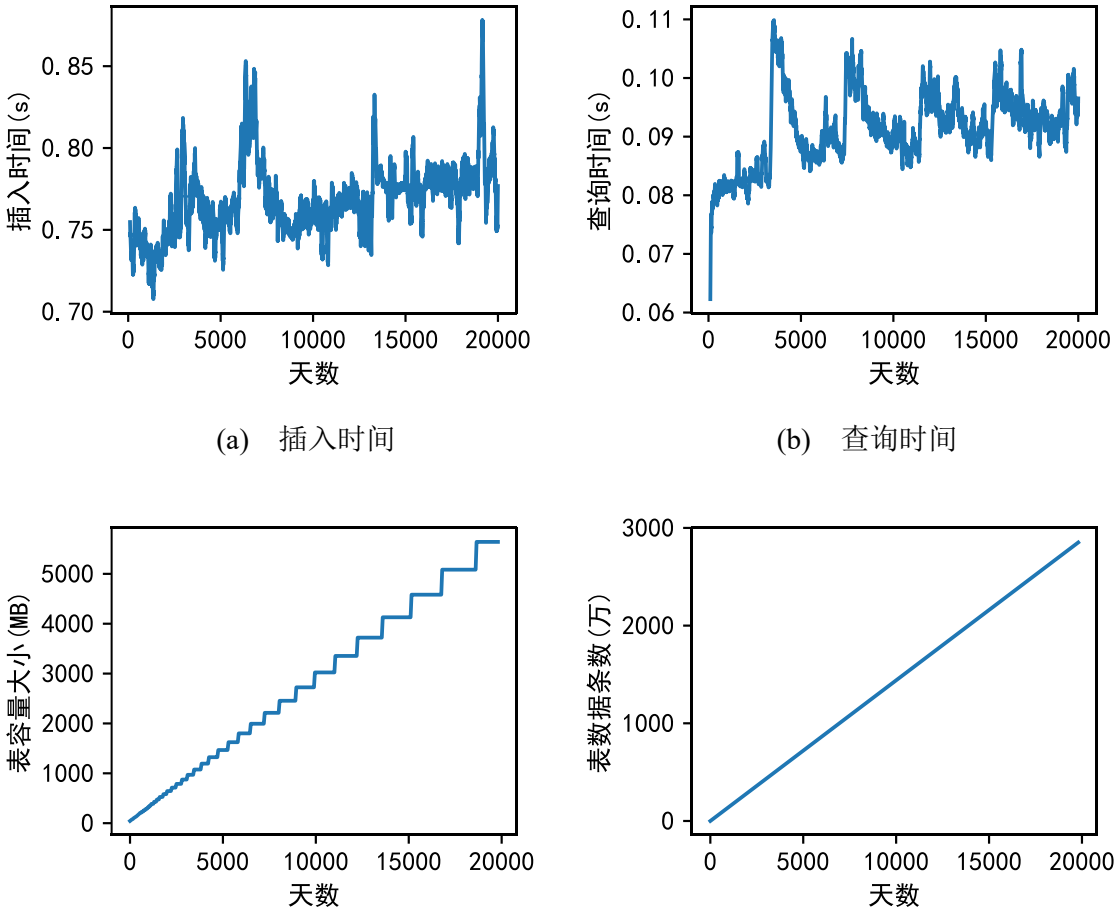
Timestamp	var	skew	kurt	abs_max	abs_mean	abs_max_top5p	abs_max_top10p	energy_sstd
1497349264	2.13748	0.0074217	-1.46878	2.75033	1.3547	2.01494	1.81693	0.172313
1497349632	0.206586	0.0067865	-0.845208	1.01816	0.383025	0.764273	0.614644	0.0140833
1497350179	0.41418	-0.00899372	-1.05685	1.36242	0.556589	0.98928	0.844901	0.0195346
1497350437	0.486783	0.0018761	-0.954814	2.60558	0.611391	1.06353	0.890652	0.048167
1497350903	0.547107	0.000183154	-1.26334	1.31667	0.652837	1.12316	0.990405	0.0269858
1497351954	1.40835	0.0142561	-0.949611	3.3316	1.07247	1.65043	1.52742	0.344377
1497352583	1.0121	-0.00617557	-1.04895	1.99806	0.823888	1.5443	1.46142	0.0445325
1497353173	4.42168	-0.00149016	-1.62962	3.3331	1.97139	2.93746	2.55495	0.22981
1497353773	3.19558	-0.00316772	-1.53549	3.01022	1.66356	2.6202	2.27782	0.0384956
1497354373	1.21105	0.00230002	-0.85713	2.86396	0.905901	1.73218	1.47455	0.102646
1497354973	3.12875	0.000463217	-1.19005	3.21535	1.57995	2.85871	2.46495	0.0256132
1497355573	2.88203	-0.000206913	-1.52992	2.66708	1.57612	2.50095	2.25794	0.0442593
1497356173	2.25805	0.00213643	-1.31863	2.77096	1.34803	2.18594	1.93206	0.111954
1497356773	2.41928	0.00257335	-1.55709	2.54558	1.45985	2.22419	2.09294	0.0955601

图 5.7　特征数据库结果

在特征数据库中使用过程中，插入和查询最多的表是按台站和信号分量分表的特征表。在实际的使用场景下，按照目前每分钟采集一条原始数据的连续采集策略，每个表每天会增加 1440 条数据，而且这些数据是顺序插入到表末尾。本文设计了一个实验来

测试随着数据量的增加，单表性能对应的变化情况。数据选取 90 号台站 2020 年 1 月 1 日真实提取的 1440 条的地声特征，并将日期修改至 1970 年 1 月 1 日。每次将这些数据加上随机值生成下一天模拟数据并插入值数据库，一共重复 20000 次。每一次插入后将随机选取有数据的 1 天进行查询，统计每次插入和查询所需时间。插入时间和查询时间以 100 为时间窗口求均值来平滑。此外每隔 50 天将查询一次表容量大小以及具体数据条数。最终结果如图 5.8 所示。

随着数据量的增加，插入和查询单日 1440 条数据的时间较为稳定，没有太大变化。其中插入时间从 0.75s 左右缓慢增加到 0.80s 左右，查询时间从 0.08s 缓慢增加到 0.1s 左右。单表容量和总共数据条数逐渐增加，其中单表容量在后期阶梯示增加。当插入 20000 天数据时，单表容量为 5640MB，总数据条数为 2880 万条。插入和查询时间稳定主要由于插入为顺序插入，而且集中添加在数据表末，查询也是基于主键的有序查询。单表的插入和查询时间符合实际的使用需求。



(a) 插入时间　　(b) 查询时间

(c) 表容量大小　　(d) 表数据条数

图 5.8 特征数据库随数据量性能的变化

5.2.5　特征数据流整体测试

完成对单个模块的测试后，本小节将整体运行并测试特征数据流。整体特征数据流将运行 4 个原始数据下载进程，2 个数据预处理进程，3 个特征提取进程和 1 个数据库接入进程。运行时将记录日志，效果如图 5.9 所示。

```
2020-05-07 20:59:25,772 - MainThread - 30022 - INFO: No file found on 20200411
2020-05-07 20:59:25,988 - MainThread - 30033 - INFO: get [240, '1', '20200411'] from buffer, size=1439, ra_extracter get 349 job, there are(is) 1 object in buffer
2020-05-07 20:59:25,989 - MainThread - 30033 - INFO: 349 data has been get, 135 left
2020-05-07 20:59:25,990 - MainThread - 30039 - INFO: get [251, '1', '20200411'] from buffer, size=0, ra_extracter get 350 job, there are(is) 1 object in buffer
2020-05-07 20:59:25,990 - MainThread - 30039 - INFO: 350 data has been get, 134 left
2020-05-07 20:59:25,990 - MainThread - 30039 - INFO: Extract Feature on [251, '1', '20200411']
2020-05-07 20:59:25,990 - MainThread - 30033 - INFO: Extract Feature on [240, '1', '20200411']
2020-05-07 20:59:25,994 - MySQL storager 2 - 29701 - INFO: Save feature on [251, '1', '20200411'] start!
2020-05-07 20:59:25,997 - MainThread - 30039 - INFO: get [251, '3', '20200411'] from buffer, size=0, ra_extracter get 351 job, there are(is) 0 object in buffer
2020-05-07 20:59:25,997 - MainThread - 30039 - INFO: 351 data has been get, 133 left
```

图 5.9　特征数据流运行效果

测试分为两部分。首先针对历史原始数据的特征提取。选择已经长期安装的台站提取所有的历史电磁和地声数据，具体选择 2017 年 6 月安装的 5 个台站，提取从安装到 2020 年 4 月 30 日的数据，结果如表 5.7 所示。总体来看，对于这些台站，每个台站累积的数据约为 100GB，其中电磁扰动的数据略大于地声的数据。每个台站提取特征大约需要 5 个小时，其中电磁扰动数据的时间长于地声数据。数据处理速度在 40.5Mb/s ~53.8Mb/s。

表 5.7　台站历史数据特征提取流运行时间

台站名	信号分量	数据条数 (条)	运行时间	原始数据大小 (GB)	数据处理速度 (Mb/s)
九寨沟防震减灾局	电磁扰动	852422	2:26:40	48.7760	45.4069
	地声	849264	2:33:48	48.5953	43.1412
金川防震减灾局	电磁扰动	870982	2:44:05	49.8380	41.4702
	地声	863733	2:46:47	49.4232	40.4610
松潘地震台	电磁扰动	766991	2:28:51	43.8876	40.2541
	地声	742625	2:18:02	42.4933	42.0300
茂县测点	电磁扰动	833870	2:29:33	47.7144	43.5601
	地声	818144	2:20:28	46.8146	45.5056
汶川防震减灾局	电磁扰动	845922	2:09:48	48.4041	50.9160
	地声	845034	2:02:39	48.3532	53.8255

另一部分针对每天新上传的原始数据，测试所需要的时间，具体选择 2020 年 4 月

11 日~2020 年 4 月 20 日 10 天的数据进行测试，结果如表 5.8 所示。可以看到，由于一些台站没有数据，一天的原始数据在 45 万条左右，对应的数据量在 25~26GB 之间。特征生成流提取一天数据需要 64~83 分钟，对应每秒钟处理的原始数据为 41.9~54.5Mb/s，目前整个特征生成流能满足每日提取新上传数据特征的需求。

表 5.8　每日新上传原始数据特征提取流运行时间

日期	电磁扰动数据 (条)	地声数据 (条)	运行时间	原始数据大小 (GB)	数据处理速度 (Mb/s)
4 月 11 日	229356	217656	1:19:42	25.5782	43.8178
4 月 12 日	229892	217465	1:04:43	25.5980	54.0042
4 月 13 日	228358	217699	1:23:06	25.5236	41.9353
4 月 14 日	226627	217592	1:20:06	25.4184	43.3266
4 月 15 日	227055	218767	1:23:27	25.5101	41.7374
4 月 16 日	232473	222947	1:14:36	26.0593	47.6940
4 月 17 日	234179	225046	1:15:54	26.2771	47.2687
4 月 18 日	233021	223824	1:12:29	26.1409	49.2403
4 月 19 日	232967	224467	1:05:34	26.1746	54.5049
4 月 20 日	234210	225729	1:16:48	26.3179	46.7874

总体来看，无论是从历史数据还是从每日新增数据提取特征数据处理速度都在 40.5~53.8Mb/s 之间，目前的速度瓶颈在于原始数据的下载速度。通过提供的接口在主服务器下载数据时，主服务器需要与从服务器网络通讯以确保数据正确性。再加以缓存更新机制，这些都降低了接口的速度。

5.3　AETA 原始数据异常风险库

5.3.1　异常检测模块

在特征数据流从原始数据提取特征并持久化存储在 MySQL 数据库，异常风险库可以从数据库中读取特征数据并进行异常提取。

选择 90 号茂县测点 2017 年 6 月 12 日~2020 年 4 月 30 日电磁原始数据中与 ULF 相关的的特征数据提取异常，结果分别如图 5.11~5.13 所示。

为了减少分钟数据的扰动，同时去除日周期的波动，首先对峰度以小时均值聚合，

并选择当天以及过去 27 天同一时刻的数据计算异常值，一天共输出 24 个异常值值。图 5.10 展示了 ULF 偏度特征在 2019 年 3 月 1 日~2019 年 10 月 1 日人为活动干扰较小的凌晨 2 点滑动四分位法检测异常值。其中蓝线为峰度特征，红线为检测到的异常值值，上方的线段表示附近的地震，每一格为 2 级。可以看到，滑动四分位法法能把偏度特征中的突变点检测出来，地震密集的区间频繁检测出异常，其中最大的突变点之后发生了一次大震。

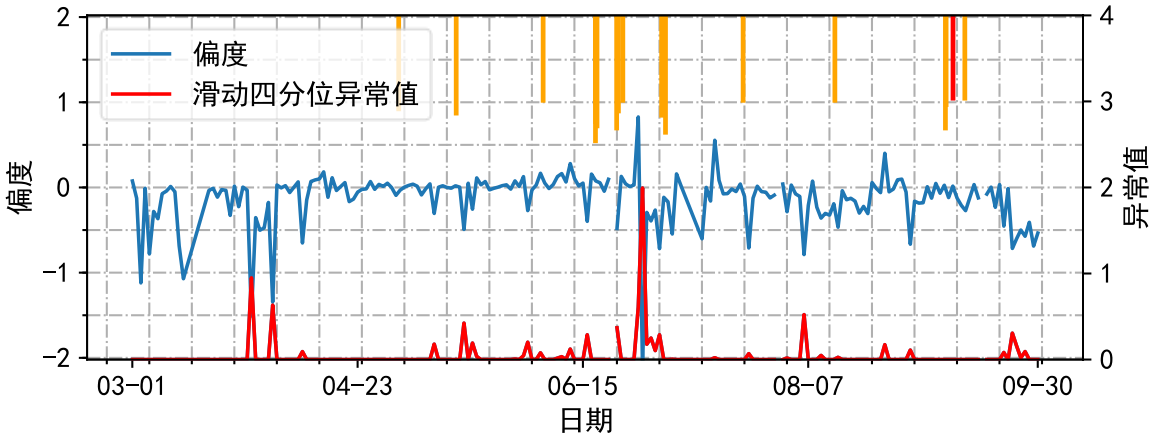


图 5.10　滑动四分位法异常检测效果

图 5.11 展示了 3sigma 的异常检测效果。所做处理处理和滑动四分位法相同。与前者相比，当 ksigma 检测中 k 取 3 时，检测出的异常值更少，但均发生在地震频繁的区间，且具有更好的映震关系。

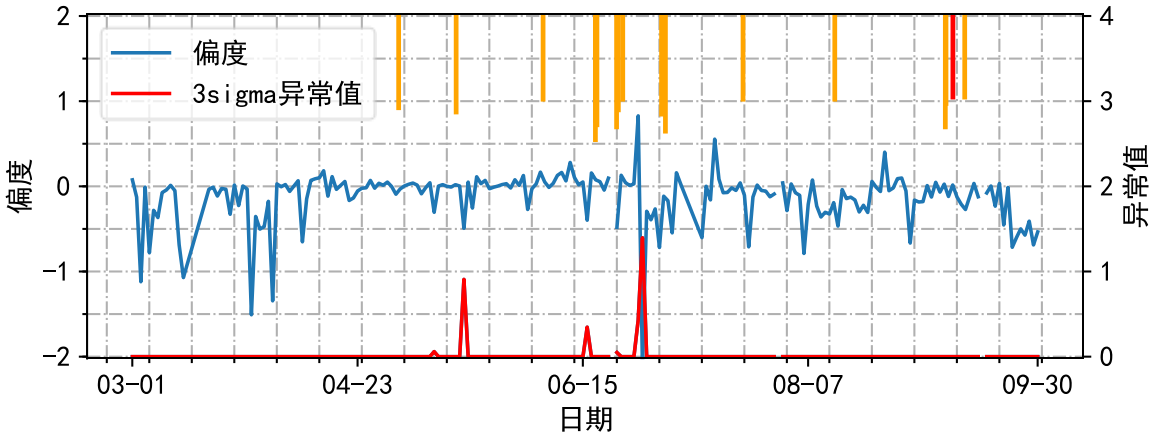


图 5.11　3sigma 异常检测效果

与前两者通过单一特征检测相比，孤立森林综合多个特征检测异常。选择按小时聚合的 ULF 特征，通过孤立森林算法判断每小时数据是否为异常值并按天统计当日异常次数，结果如图 5.12 所示。可以看到，在地震密集区间孤立森林频繁检测出异常，且异常出现最多的一天随即发生了红色柱子表示的台站 100km 以内 3 级以上的地震。

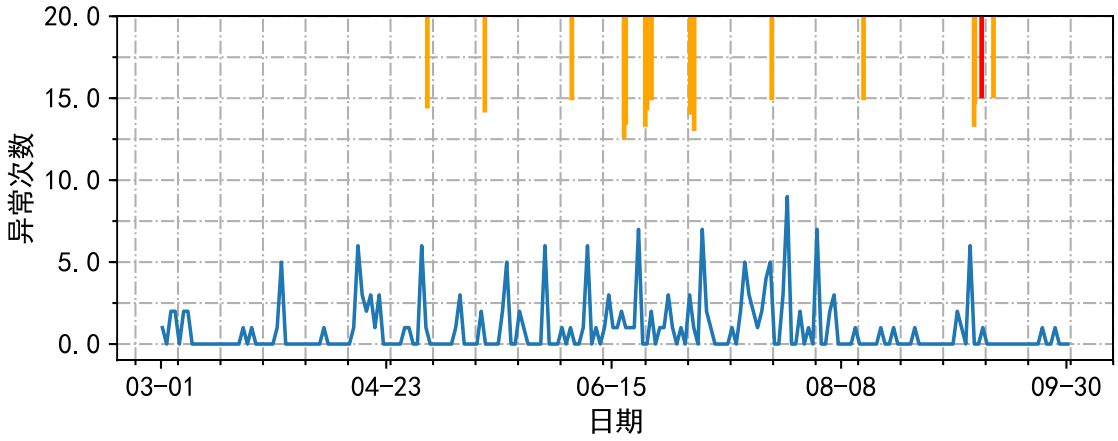


图 5.12　孤立森林异常检测效果

5.3.2　异常评价模块

本小节将以上一小节针对 90 号茂县测点检测到的异常为例子，展示异常评价模块的效果。对于 ULF 偏度特征通过 IQR 检测的异常，图 5.13~5.15 展示了三种分析得到与地震的相关性。

图 5.14 的方差分析显示，检测出的异常还没达到显著性水平，但反映出和地震标签一定的相关性。

	F	PR(>F)
abnormal	2.110284	0.146667

图 5.13　方差分析结果

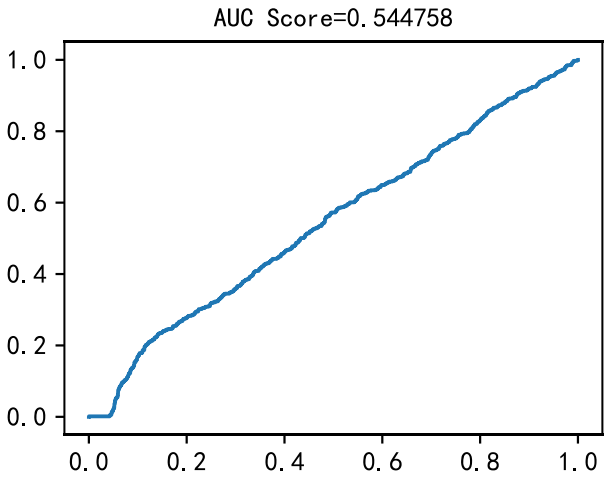


图 5.14　AUC 指标得分

图 5.15 展示了异常指标的 ROC 曲线以及对应的 AUC 得分。AUC 得分为 0.5447，略高于随机的 0.5，但尚未体现出明显的相关性。

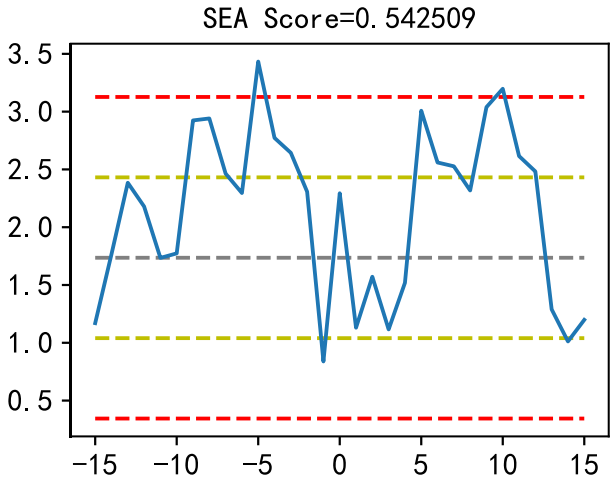


图 5.15　SEA 分析结果

图 5.16 展示了 SEA 分析的效果。其中灰色线为均值μ，黄线为$\mu \pm \sigma$，红线为$\mu \pm 2\sigma$。可以看到在发震日附近异常较低，而在发震前后异常较高，其中在地震发生前 5 天，SEA 的值超过了$\mu \pm 2\sigma$，表现出和地震较强的相关性。

异常评价模块现在集成的 3 种评价指标从不同角度描述了异常风险指标和地震的相关性，反映了异常的风险程度。此外，可以基于本框架添加新的异常评价指标，并使用数据库接口配置数据库，从而集成更多的异常评价方法。

5.3.3　异常风险库

异常风险库主要持久化存储异常检测模块和异常评价模块得到的结果，并根据其他模块需要，返回对应的异常值和异常与地震相关程度。本小节主要对异常风险库进行展示。

StationID	abnormalID	time	sum	max	maxHour	count
19	1001	2018-11-04 00:00:00	0.28668	0.20527	15	2
19	1001	2018-11-05 00:00:00	2.0398	0.68129	7	6
19	1001	2018-11-06 00:00:00	0.12942	0.12942	19	1
19	1001	2018-11-09 00:00:00	0.0061	0.0061	8	1
19	1001	2018-11-21 00:00:00	0.11323	0.11323	23	1
19	1001	2018-12-29 00:00:00	0.43661	0.36607	19	2
19	1001	2019-01-01 00:00:00	0.10399	0.10399	23	1
19	1001	2019-01-18 00:00:00	0.26693	0.196	22	2
19	1001	2019-01-19 00:00:00	1.9633	0.69666	17	3
19	1001	2019-01-20 00:00:00	0.37546	0.21525	18	2

图 5.16　电磁异常风险指标记录

选择四川地区 2018 年 10 月 1 日~2020 年 3 月 30 日的电磁特征数据进行异常提取

和评价，最终存入数据库。图 5.16 展示了入库的异常风险记录。为了减少数据量，入库的异常风险将按天聚合，通过异常风险得分总和，异常风险最大值，最大值出现的小时以及当天出现的异常风险次数来描述。另外，对于没有出现异常风险的天数，将不被录入数据库。

图 5.17 展示了存储的异常风险指标得分。异常风险库将记录每个台站，某种特征在某种异常评价方法和打标函数下的相关性指标，包括方差分析，AUC 指标和 SEA 分析。startDate 和 endDate 则为选择评价的时间区间。

abnormalID	StationID	labelID	p_value	auc	seaScore	startDate	endDate
1001	19	1	0.220512	0.438889	2.31044	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	24	1	0.241471	0.584071	2.39427	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	33	1	0.194397	0.456621	0.403327	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	34	1	0.664608	0.496988	1.14344	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	39	1	0.430043	0.634921	4.62844	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	43	1	0.043185	0.534816	0.616193	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	55	1	0.613356	0.289916	0	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	73	1	0.415015	0.414286	1.77672	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	75	1	0.143847	0.422151	2.27634	2018-10-01 00:00:00	2020-03-30 00:00:00
1001	77	1	0.545285	0.467593	5.13504	2018-10-01 00:00:00	2020-03-30 00:00:00

图 5.17　电磁异常风险指标得分

5.3.4　展示层

展示层使用交互式可视化框架 Dash 进行可视化，可以通过 web 访问。展示层有 4 种展示方式，包括波形图、数据分布图、频谱热力图和时空分布图。

波形图的界面如图 5.18 所示。最上方的选项卡是共有的界面，可以通过点击跳转到对应的页面。波形图可以选择展示单台站单特征，或者同时展示同一台站多个特征或多个台站同一特征。当选择单台站单特征时，可以点击同期对比按钮，对比该特征和过去的区别。在天模式下为过去一天和过去一周的数据。在周模式下为过去一周和过去 4 周的数据。在月模式下为过去一个月的过去一年的数据。图 5.18 为 121 号九寨沟防震减灾局 2019 年 5 月 29 日的特征数据和过去的对比，可以移动鼠标查看每个点当日和之前的对比。

由于数据点较多，波形图可以选择是否自动聚合。为了日常分析的需要，是否显示地震也是其中一个选项。此外，网页还对异常输入做提示，如没有数据台站名，信号分量或特征。各个选项框支持搜索功能。

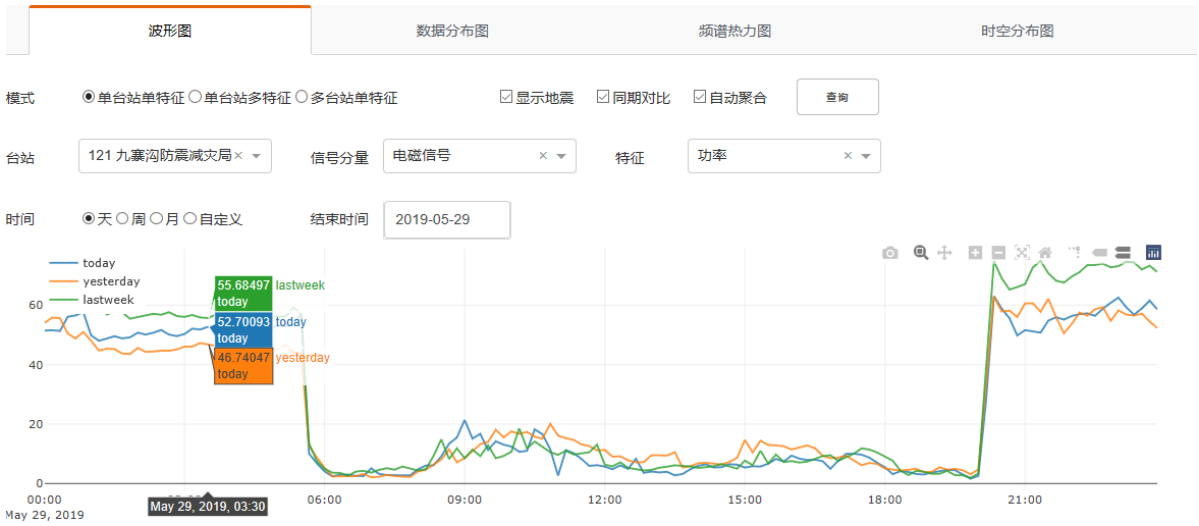


图 5.18　波形图界面

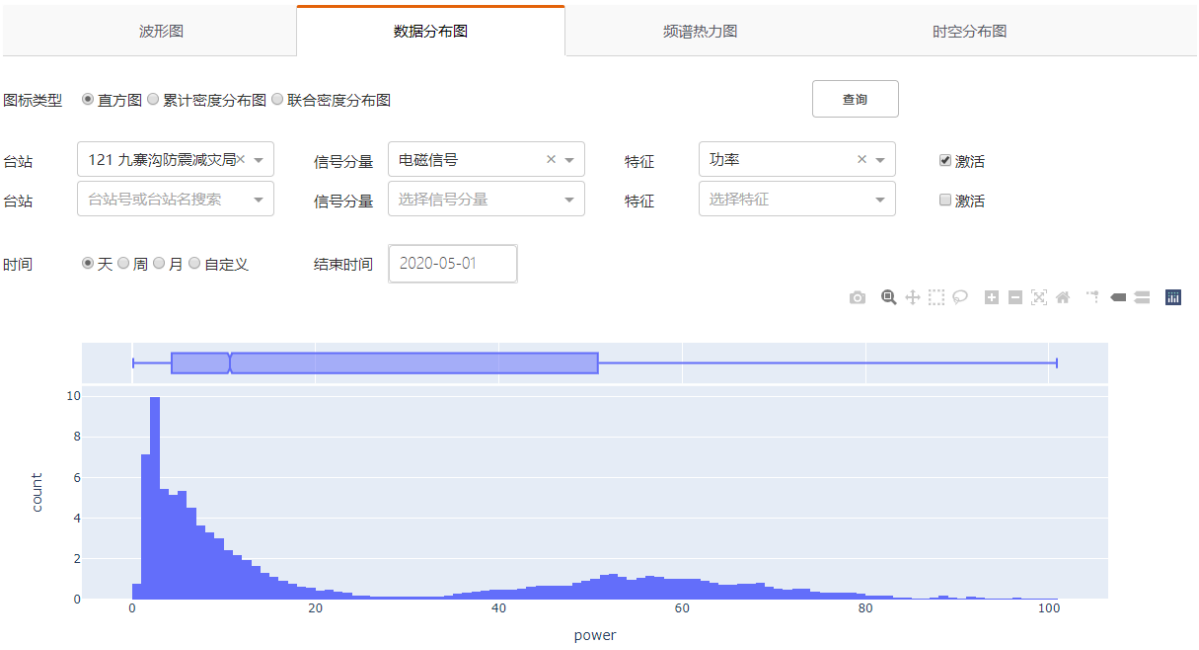


图 5.19　数据分布图界面

数据分布图如图 5.19 所示，可以选择直方图，密度分布图和联合密度分布图。数据分布图能让使用者对数据有更直观的认知。数据分布图的界面和使用方法和波形图相似，同时也进行了对应的异常处理。

图 5.20 展示了频谱热力图的界面，频谱热力图展示了不同频段能量随着时间的变化。频谱热力图也有选择聚合的模式，同时为了查看不同频段能量变化分布的情况，可以选择归一化按钮。

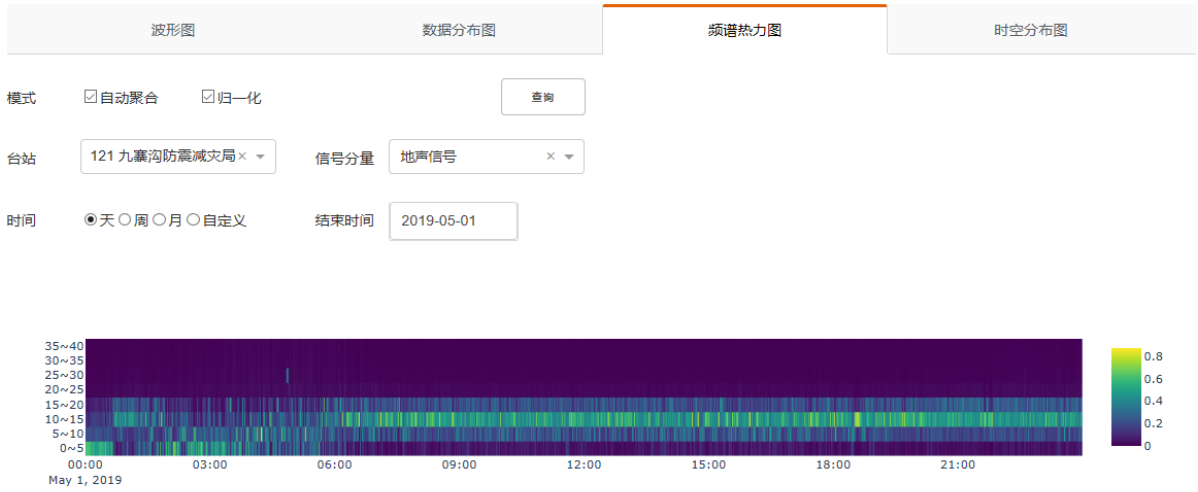


图 5.20　频谱热力图界面

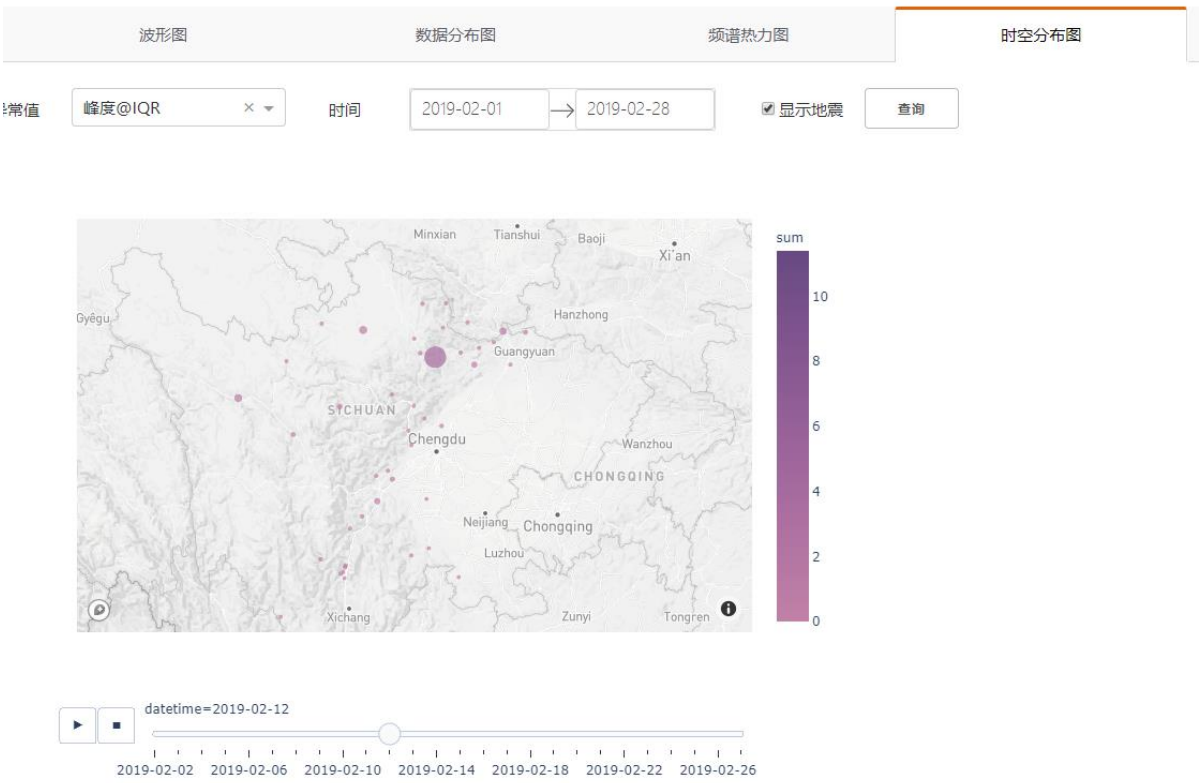


图 5.21　时空分布图界面

图 5.21 展示了时空分布图的界面。时空分布图将展示某种异常选定日期在空间中的分布，点的大小反映异常的大小。当设置好需要查询的日期范围后，可以点击下方横轴的开始按钮，自动播放异常在空间上随时间的变化。也可以手动拖放选择具体的某一天查看。

5.4　本章小结

本章针对原始数据分析平台的特征生成流和异常风险库，从功能和性能对里面的子模块和整体进行测试，最终展示了展示层可视化的结果。结果表明，数据分析平台能准确稳定地在合理时间内从大数据量的原始数据中提取特征进行持久化存储，并能针对提取的特征检测异常风险和评价异常风险与地震的相关性。最终展示层能通过 Dash 可视化框架实现交互式可视化。此外，对短时能量的优化算法能分别相对遍历法和矩阵法减少 87.87%~89.24%和 27.31%~68.41%的运行时间，短时过零率使用矢量优化后减少 99.82%的时间，从而使总体特征提取模块运行每日时间减少至 17.83s。总体来看，AETA 原始数据分析平台各层次设计合理，整体功能完善，达到了顶层设计的要求。

第六章　总结与展望

6.1　总结

中国历来是地震频发的国家，地震高活动频率，广分布范围，强破坏性的特点导致我国成为地球上受地震灾害影响最为严重的国家之一。为了对地震这世界难题进行探索并准确有效预测地震三要素，北京大学集成微系统实验室研发了集成地震前兆信号采集传输和分析的 AETA 系统。从 16 年底稳定运行至今的 AETA 系统已经积累了 TB 级别的原始数据，并从均值、振铃计数、峰值频率，峰值幅度四个方面提取了 GB 级别的特征数据。

针对现有的特征数据，经过数年的研究，目前已经取得了丰富的成果。但由于有效信息过少，提取特征前缺乏数据预处理以及部分特征数据设置不够合理，现有特征在研究复杂多变的地震仍有不足。为了充分利用 TB 级别的原始数据，从多维度挖掘特征以研究与地震之间的相关性，本文从原始数据出发提取特征，进而检测和评价异常并最终搭建数据分析平台，主要工作和创新点如下：

1、分析原始数据的特点，并根据原始数据的特点完成原始数据分析平台顶层设计。本文探究了电磁扰动和地声信号的独有形态和特点，以及共有的差异性大，噪声相对较大和数据量大的三大特点。基于需要重新挖掘数据内在特征的需求，以及原始数据量大导致生成代价大的特点，需要构建特征生成流来高速有效地提取特征。基于原始数据时空差异性和噪声相对较大的特点，研究应更关注某一个台站和自身相比的异常，并联合多个台站进行分析，需要构建异常风险库。因此，原始数据分析平台由特征生成流和异常风险库两大部分组成。本文还对每个组成部分进行详细设计，并确立了持久化存储层、接入层、数据处理层和展示层共四层架构。

2、设计与实现了特征生成流。特征生成流在运行初始阶段将按照一个台站一个分量一天的数据来划分任务，并放到任务池中。对于每一个任务，原始数据将通过原始数据下载器获取，接下来由数据预处理模块识别并处理断电重启异常、数据缺失异常和脉冲型异常。预处理后的数据将在特征提取模块中从时域、频域和小波变换出发提取特征。其中，电磁扰动信号提取 49 个特征，地声信号提取 42 个特征。最后，特征数据将通过特征数据库接入模块存入特征数据库。

3、设计与实现了异常风险库。首先构建异常检测框架，从特征数据库中获取特征数据，并使用项目组已经取得较好效果的低复杂度算法进行异常检测。异常检测算法包括基于单特征滑动四分位法和 ksigma，以及基于多特征集成学习的孤立森林算法，

检测出的异常风险指标将通过方差分析，AUC 指标和时间叠加分析三个方面评价异常和地震的相关性，最终将结果持久化存储在异常风险库中。

4、设计用于持久化存储特征和异常风险的数据库，并开发对应的数据库接口。基于存储数据的特点和访问的需求，设计了对应的特征数据库和异常风险库表的结构。对于数据量大的表格，基于实际业务的场景，选择合适的指标水平分表。

5、针对运行时间长的函数设计算法加速。特征提取模块中短时能量和短时过零率特征运行时间较长。在短时能量特征方面，经过优化和遍历法以及矩阵法相比，分别减少 87.87%~89.24%和 27.31%~68.41%的运行时间。短时过零率使用矢量优化后减少 99.82%的时间，从而使单个特征提取模块进程提取单任务的时间减少至 17.83s。

6、基于开源的 python 可视化框架 Dash 构建展示层。通过网页端，可以从波形图，数据分布图，频谱热力图以及时空图可视化查看特征数据以及其异常和分布。每个部分可以选择台站，信号分量，时间等参数，选择完成后点击查询生成图像。图像可以拖拽缩放筛选保存，当鼠标悬停数据点时可以展示信息，实现交互式可视化查询，从而让使用者对数据有更深的理解。

6.2　展望

原始数据分析平台能对海量历史及新增的原始数据提取特征，检测和评价异常，最终通过 Dash 在网页上可视化展示，但仍存在以下不足和等改进之处：

1、目前特征数据流的速度受限于原始数据下载的速度，目前的接口为合作伙伴提供查询接口，实际使用中发现无论是硬盘活动时间，对外网络带宽还是 CPU 使用率尚未得到充分利用，其性能主要受限于主从服务器的大量通信。未来需要从更底层进行开发，提高原始数据接口的下载速度，从而加快整体特征数据流的速度。

2、异常检测模块和异常评价模块目前只集成了少量算法，在未来随着研究的深入，需要添加更多有好效果的算法。同时，为了准确发现和研究异常，对于选定的异常与模型，需要每日定时生成报表与图片，方便研究人员对地震三要素的预测与研究。

3、在展示层方面，由于当前主要用于内部数据分析使用，现版本没有用户管理和个性化的设置，直接通过点击不同选项卡查看四种图片。未来需要添加用户登录，用户权限管理等功能，同时用户可以在仪表盘中自定义感兴趣的图标，在每次登录后首页展示，更高速地对数据进行分析。

参考文献

[1] 陈运泰. 地震预测:回顾与展望[J]. 中国科学:, 2009(12): 1633–1658.

[2] 徐纪人, 赵志新. 汶川 8.0 级大地震震源机制与构造运动特征[J]. 中国地质, 2010(04):135-145.

[3] 张国民. 我国的地震灾害和震灾预防[J]. 科学对社会的影响, 1999, (02): 44-47.

[4] 林科, 王新安, 张兴等. 一种适用于大地震临震预测的地声监测系统[J]. 华南地震, 2013, 33(04): 54-62.

[5] 王新安, 雍珊珊, 黄继攀,等. 基于 AETA 监测数据的地震预测研究[J]. 北京大学学报(自然科学版), 2019, 55(02):16-21.

[6] 马文娟, 刘坚, 蔡寅等. 大数据时代基于物联网和云计算的地震信息化研究[J]. 地球物理学进展, 2018, 33(2): 835–841.

[7] Jousset P, Reinsch T, Ryberg T, et al. Dynamic strain determination using fibre-optic cables allows imaging of seismological and structural features[J]. Nature Communications, 2018, 9(1): 2509-.

[8] Pisco M, Bruno F A, Galluzzo D, et al. Opto-mechanical lab-on-fibre seismic sensors detected the Norcia earthquake[J]. Scientific Reports, 2018, 8(1): 6680.

[9] Marra G, C C, R L, et al. Ultrastable laser interferometry for earthquake detection with terrestrial and submarine cables[J]. Science, 2018, 361(6401): 486.

[10] 陈运泰, 吴忠良. 国际地震学与工程地震学手册[J]. 地震学报, 2004, 26(1):110-111.

[11] 王新安, 雍珊珊, 徐伯星,等. 多分量地震监测系统 AETA 的研究与实现[J]. 北京大学学报:自然科学版, 2018,54(03):487-494.

[12] 邱泽华, 张宝红. 我国钻孔应力-应变地震前兆监测台网的现状[J]. 国际地震动态, 2002(06):6-10.

[13] 邱泽华, 石耀霖. 国外钻孔应变观测的发展现状[J]. 地震学报, 2004, 000(0S1).

[14] Xiangzeng, Kong, Nan, et al. Relationship of Stress Changes and Anomalies in OLR Data of the Wenchuan and Lushan Earthquakes[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2018, 11(8):p.2966-2976.

[15] 池顺良. 日本 9 级大震前我国钻孔应变网测到两起地块强烈受压事件[J]. 地球物理学进展, 2011, 26(5):1583-1587.

[16] 刘君, 安张辉, 范莹莹. 芦山 Ms7.0 与岷县漳县 Ms6.6 地震前舟曲台电磁扰动异常变化[C]// 2015 中国地球科学联合学术年会论文集（十）——专题 28 电磁地球物理学研究应用及其新进展、专题 29 盆地动力学与能源、专题 30 活动断层、地震构造与深部结构. 2015:15-17.

[17] Q. Ma, G. Fang, W. Li,等. Electromagnetic anomalies before the 2013 Lushan MS7.0 earthquake[J]. Acta Seismologica Sinica, 2013, 35(5):717-730.

[18] 李军辉, 李琪, 王行舟,等. 关于汶川 8.0 级地震前电磁扰动异常的讨论[J]. 华南地震, 2011, 031(004):76-84.

[19] Bagiya M S, Sunil A S, Sunil P S, et al. Efficiency of coseismic ionospheric perturbations in identifying crustal deformation pattern: Case study based on Mw 7.3 May Nepal 2015 earthquake[J]. Journal of Geophysical Research Space Physics, 2017, 122(6).

[20] Guo J, Li W, Yu H, et al. Impending ionospheric anomaly preceding the Iquique Mw8.2 earthquake in Chile on 2014 April 1[J]. Geophysical Journal International, 2015, 203(3):1461-1470.

[21] Rui, Yan, Michel,等. Statistical Study on Variations of the Ionospheric Ion Density Observed by DEMETER and Related to Seismic Activities[J]. Journal of Geophysical Research Space Physics, 2017,122(12).

[22] 张文蕾, 杨奕, 翁骋. 地震地下流体实时监测与地震预测研究[J]. 科技创新导报, 2018, v.15；No.436(04):28-29.

[23] 王广才, 沈照理. 地震地下水动态监测与地震预测[J]. 自然杂志, 2010(02):32-35.

[24] 王志敏. 地震地下水动态监测与地震预测研究[J]. 科技经济导刊, 2016, 000(003):P.112-113.

[25] Bleier T, Freund F. Earthquake [earthquake warning systems][J]. IEEE Spectrum, 2005, 42(12):p.22-27.

[26] 潘黎黎, 曾佐勋, 王杰. 芦山地震(M_S7.0)及玉树地震(M_S5.2)震前次声波异常信号分析[J].地学前缘,2013,20(06):73-79.

[27] 夏雅琴, 崔晓艳, 李均之等. 震前次声波异常信号的研究[J]. 北京工业大学学报, 2011(3): 463–469.

[28] Kristoffer Walker, Alexis Le Pichon, Tae Sung Kim, et al. An Analysis of Ground Shaking and Transmission Loss From Infrasound Generated by the 2011 Tohoku Earthquake[J]. Journal of Geophysical Research Atmospheres, 2013, 118(23):12-12,851.

[29] 王新安, 雍珊珊, 黄继攀,等. 基于 AETA 监测数据的地震预测研究[J]. 北京大学学报(自然科学版), 2019, 55(02):16-21.

[30] 吕亚轩,王新安,黄继攀,雍珊珊.基于 AETA 电磁扰动对九寨沟 Ms 7.0 级地震的研究[J].北京大学学报(自然科学版),2019,55(06):1007-1013.

[31] 张继艳, 王新安, 雍珊珊,等. 基于 ARIMA 模型的九寨沟 7.0 级地震前兆异常检测[J]. 华北地震科学, 2019, 37(01):31-36.

[32] 李柏杭, 王新安, 雍珊珊,等. 人工免疫算法在 AETA 异常检测中的应用研究[J]. 计算机技术与发展, 2019, 29(03):7-11.

[33] 刘高川, 李正媛, 王建国,等. 地震前兆台网数据跟踪分析平台设计简[J]. 大地测量与地球动力学, 2016(9):841-846.

[34] 彭俊芳, 郑贵洲. 广西地震工程分析信息系统平台属性数据库的设计与构建[J]. 地矿测绘, 2017, v.33;No.129(04):18-20+48.

[35] 康凯. 基于 Hadoop 架构的地震大数据平台研究与实现[D].中国地震局地震研究所,2016.

[36] 燕云, 卢山, 刘天龙,等. 辽宁省地震前兆应急监控与数据处理平台的 开发与应用[J]. 防灾减灾学报, 2019(2).

[37] Nikolaou A S. A GIS Platform for Earthquake Risk Analysis[D]. State University of New York at Buffalo, 1998.

[38] T. C. Vance, N. Merati, S. M. Mesick,等. GeoModeler: Tightly linking spatially-explicit models and data with a GIS for analysis and geovisualization[C]// Acm International Symposium on Advances in Geographic Information Systems. ACM, 2007.

[39] Hardisty F, Robinson A C. The geoviz toolkit: using component-oriented coordination methods for geographic visualization and analysis[J]. International Journal of Geographical Information ence, 2011, 25(1-2):p.191-210.

[40] Levente J Klein, Fernando J Marianno, Conrad M Albrecht,等. PAIRS: A scalable geo-spatial data analytics platform[C]// IEEE International Conference on Big Data. IEEE Computer Society, 2015.

[41] 刘晨光, 王新安, 雍珊珊等. AETA 多分量地震监测系统的数据存储与安全系统[J]. 计算机技术与发展, 2018, 28(12): 7-12.

[42] 林珠, 邢延. 数据挖掘中适用于分类的时序数据特征提取方法[J]. 计算机系统应用, 2012, 21(10): 224-229.

[43] 刘慧, 谢洪波, 和卫星,等. 基于模糊熵的脑电睡眠分期特征提取与分类[J]. 数据采集与处理, 2010(04):70-75.

[44] Greene B R, Boylan G B, Reilly R B, et al. Combination of EEG and ECG for improved automatic neonatal seizure detection[J]. Clinical Neurophysiology, 2007, 118(6):1348-1359.

[45] Gao G Q. Computerised detection and classification of five cardiac conditions[D]. Auckland University of Technology, 2003.

[46] 钱争鸣. Application of ARCH Group Measurement Models in the Study of Financial Market [J]. 厦门大学学报(哲学社会科学版), 2000, 000(003):126-129.

[47] Mrchen F. Time series feature extraction for data mining using DWT and DFT[C]. Technical Report No.33, 2003.

[48] Hossan M A, Memon S, Gregory M A. A novel approach for MFCC feature extraction[C]// Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on. IEEE, 2011.

[49] 高瑞华. 多种预处理方法在语音检测中应用效果的比较研究[D]. 浙江工业大学, 2004.

[50] 王春. 基于小波和分形理论的齿轮故障特征提取及噪声的和谐化研究[D]. 重庆大学, 2006.

[51] 金解放,赵奎,王晓军,赵康.岩石声发射信号处理小波基选择的研究[J].矿业研究与开发, 2007(02):12-15.

[52] 董胡. 基于窗函数与 MATLAB 的数字 FIR 滤波器设计[J]. 微型电脑应用,2016, 32(3): 30-32.

[53] Liu F T, Ting K M, Zhou Z H. Isolation Forest[C]// Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on. IEEE, 2009.

[54] Khandelwal I, Satija U, Adhikari R . Efficient financial time series forecasting model using DWT decomposition[C]// IEEE International Conference on Electronics. IEEE, 2015.

[55] Madan R, Sarathimangipudi P. [IEEE 2018 Eleventh International Conference on Contemporary Computing (IC3) - Noida, India (2018.8.2-2018.8.4)] 2018 Eleventh International Conference on Contemporary Computing (IC3) - Predicting Computer Network Traffic: A Time Series Forecasting Approach Using DWT, ARIMA and RNN[C]// 2018:1-5.

[56] Baum L E, Petrie T, Soules G, et al. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains[J]. Annals of Mathematical Statistics, 1970, 41(1):164-171.

[57] 李和平, 胡占义, 吴毅红,等. 基于半监督学习的行为建模与异常检测[J]. 软件学报, 2007, 018(003):527-537.

[58] HOANG X A, HU J. An efficient hidden Markov model training scheme for anomaly intrusion detection of server applications based on system calls[C]// Proceeding of IEEE International Conference on Networks. Singapore:IEEE Computer Society, 2004:470-474.

[59] 王琼, 倪桂强, 潘志松,等. 基于改进隐马尔可夫模型的系统调用异常检测[J]. 数据采集与处理, 2009(04):112-117.

[60] Forrest S, Perelson A S, Allen L, et al. Self-nonself discrimination in a computer[C]// IEEE Computer Society Symposium on Research in Security & Privacy. IEEE, 2002.

[61] GONZALEZ F A. A study of artificial immune systems applied to anomaly detection[D]. Tennessee :The University of Memphis, 2003.

[62] 董永贵, 孙照焱, 贾惠波. 时间序列中异常值检测的负向选择算法[J]. 机械工程学报, 2004, 040(010):30-34.

[63] 汪慧敏, WANGHui-min. 基于改进负选择算法的异常检测[J]. 计算机技术与发展, 2009, 19(8):41-44.

[64] Gong M, Zhang J, Ma J, et al. An efficient negative selection algorithm with further training for anomaly detection[J]. Knowledge Based Systems, 2012, 30:p.185-191.

[65] Han, Peng, Hattori, Katsumi, Hirokawa, Maiko, et al. Statistical analysis of ULF seismomagnetic phenomena at Kakioka, Japan, during 2001-2010[J]. Journal of Geophysical Research Space Physics, 119(6):4998-5011.

[66] Prager M H, Hoenig J M. Superposed Epoch Analysis: A Randomization Test of Environmental Effects on Recruitment with Application to Chub Mackerel[J]. Transactions of the American Fisheries Society, 1989, 118(6):608-618.

[67] S. E. Milan, A. Grocott, B. Hubert. A superposed epoch analysis of auroral evolution during substorms: Local time of onset region[J]. Journal of Geophysical Research Atmospheres, 2010, 115(A5):-.

[68] Grocott A, Wild J A, Milan S E, et al. Superposed epoch analysis of the ionospheric convection evolution during substorms: onset latitude dependence[J]. Annales Geophysicae, 2009, 27(2): 591-600.

攻读硕士学位期间的科研成果

已录用文章：

[1] **Binyan Ma**, Shanshan Yong, Xinan Wang. A Fast SNR-based Vibration Events Detection Algorithm for AETA Geo-acoustic Data. IEEE 5th Information Technology and Mechatronics Engineering Conference[C], 2020.(EI)

已公开专利：

[1] **马滨延**, 王新安, 雍珊珊, 张兴, 黄继攀, 冯远豪等. 用于山体滑坡的监测数据处理方法和山体滑坡预报方法：中国, 201810871963.7[P]. 2018-08-02.

[2] 王新安, 雍珊珊, 张兴, 何春舅, **马滨延**. 一种断裂带的活动监测方法、勘探方法和装置：中国，201810589885.1[P]. 2018-06-08.

致谢

漫无止境的一月寒假终于看到了尽头，加里敦大学毕业只流传在笑谈间，研究生阶段也步入了尾声。2020 年注定是被载入史册的一年，新年前后戈恩和苏莱曼尼事件闹得沸沸扬扬，回首一看才发现这只是一群盘旋的黑天鹅中最不起眼的几只。世界发生巨变的转折点往往是那么地悄无声息，但从英国脱欧扇起的飓风已成为房间里的大象。

无论外面如何风起云涌，拥有较为独立的环境的校园屏蔽给了风雨提供了宁静的氛围，让我们在感受着时代潮水起伏的同时也能把主要的精力投在科研。四年前参加夏令营的场景仿佛就在昨天，那时的校园到处遍布着装修的痕迹，4 米高的蓝铁皮将教学区团团包裹，就像披着婚纱一样充满着神秘感。如今，学校的各个角落已经印在我脑海中，里面都是星星点点快乐的回忆，感谢学校提供这桃花源般的世界。

我和学校以及项目组一切的缘起还是那大三的保研分享会上，通过这次分享会，我结识了即将加入王老师整个大实验室的肖康林师兄。他详尽地给我进行介绍了校园和实验室的情况，并客观诚恳给我提供建议，在日后的生活中也给我提供了诸多帮助，感谢已经成功转博的肖博士成为我梦想的引路人。

在几个月后的夏令营期间，第一次踏上这片土地，在这里我见到了未来的导师王新安老师。在等待面试排队时，我不停在头脑中幻想着面试的场景，万万没料到，等待我的是那充满中国古典风的办公室和那和蔼可亲的王老师。除了常规的面试之外就票房这个你并不熟悉的领域畅聊，透露出敏锐的洞察力。在日后的课程和开会中，您那对未来行业发展趋势的高远视野，解决极富实际意义世界难题的家国情怀给我留下很深印象，在这里我要对王新安导师表示最诚挚的谢意。

时间转到了大四下学期，我怀着惴惴不安的心情来到这崭新的城市开始一段较长时间的生活，来到实验室完成本科毕设。我还清楚地记得当年研三的庞瑞涛师兄帮我开实验室的门，介绍我和项目组其他成员共同午餐，并在之后带我购置所需的日常用品，项目组共同午餐的传统就这样从你们那届传承下来。当天下午，项目组组长雍珊珊博士拉上黄继攀博士了解我的情况，并指导我本科毕设。在这几年来，虽然作为项目组组长且与其他成员都隔了好几级，师姐却是那么亲切，关注每个成员的身心成长和科研进度，和整个项目组融为一体，大家都以珊珊姐称呼。在工作中严谨认真且凡事亲力亲为，硬件出身的她为了适应项目组工作重心的转移努力学习最新的知识，并很好地掌控者研究的节奏。而广泛阅读地震领域最新的研究成果的黄博对项目有很深的理解，在数据分析中总能给我们提供思路，启发我们可研究的方向。

在我进行本科毕业设计时结识了各位师兄师姐。其中 18 级的刘晨光师兄是那一届

的独苗。在我研一的阶段，整个项目组都靠光哥撑着，上到整个系统的架构，下到日常运维的琐事你是我们信心的保证。另外，很幸运能在实验室遇到同为石门中学毕业的冯远豪师兄，在未来的日子里师兄给我科研和生活给予了方方面面的指导，可以说我的研究生阶段都在追随者师兄的脚步，而师兄的为人处世之道更是我所仰慕和钦佩的。另外感谢李柏杭、周康生师兄，以及吕亚轩、张继艳师姐，是你们在数据分析打下的基石指引我们稳步前进。

告别了本科的学习生活，我在数月后再一次踏入了这片土地。项目组有安装设备来加深对项目理解的传统，我们都期望着能有一天能外出安装，没想到能在研一上学期就能实现这个梦想。徐伯星老师带我们出去安装，特意让第一次安装的我们自行摸索并最后点评。虽然过程磕磕绊绊意外不断，这样的教育方式却让我们有最多的收获。

在更深入了解实验室，生活步入正轨后，我才发现科研生活的背后还有很多老师在背后默默付出。王培老师在对外联系和日常运维方面确保着 AETA 系统的稳定进行。何春舅老师则一手操办财务报销和项目管理的工作，在这毕业季作为答辩秘书为我们忙前忙后。陈红英老师则从工作环境和设备方面将实验室管理地井井有条，感谢他们在台后的付出。

日常交集最多的还是同届的战友们，我们一同在开学破冰营中欢笑，为完成课程或赶项目进度而熬夜，为调研整理地震领域最新进展而合作，为求职而互相交流信息和帮扶。在三年间我们共同经历了研究生各个阶段，一起成长并建立了深厚的感情。如今即将离别，祝吕孟轩、杨兴文、刘聪、丘志成、李丹各位同学前程似锦，走出精彩的人生。

时间匆匆而过，转眼间已升到研二，杨超、王晶师弟以及蒋冰慧、郭琴梦师妹加入了项目组。看着师弟师妹们为熟悉项目而摸索的样子，仿佛看到了过去的自己。又是一年飞逝，项目组又迎来了鲍振宇、刘一宾、马一中、谢锦汉和张馨宝师弟。此时的研二已经可以独挡一面，在做出研究成果的同时也能对研一的同学进行指导，祝福师弟师妹们能在未来取得更多的进展。

如今即将毕业，在过去实验室的生活中，最能感受到项目组历史的传承是聚会的时刻。韩朝相、曾敬武、金秀如、庞瑞涛等师兄总能用他们的人生经验给我们启迪，甚至李柏杭师兄不远千里从新加坡回来探访。感谢他们传承着项目组精神，我也会把这份精神继续传下去。

最后感谢我的家人，感谢多年来他们对我的理解和支持，使他们给予了我一个温馨而自由的家，让我走到今天。

虽然世界也难回到过去，但新世界的机遇才刚刚开始！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日