北京大学

硕士研究生学位论文

题目：**AETA 数据的压缩处理与地声特征的提取和分析**

姓　　名：	吕孟轩
学　　号：	1701213534
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

对地震三要素的预测是世界公认的难题。尤其是对地震短临预测研究的进展，远远无法满足保护人民生命和财产安全的基本需求。在这个背景下，北京大学深圳研究生院集成微系统重点实验室研制了多分量地震监测系统 AETA。该系统包括可以采集地震数据的地声和电磁探头传感器、处理地震数据的地面数据处理终端、云服务器以及数据存储设备。可以全方位地服务于地震前兆异常的监测、地震三要素的预测以及断裂带的测量等科研工作。本文主要围绕 AETA 多分量地震监测系统的数据处理和数据分析环节，做出了如下工作：

1. 提出了对 AETA 数据采集流程的改进方案，将数据的采集方式从抽样采集改进为连续采集方式，提高了数据的连续性，在优化了 AETA 数据映震效果的同时，也满足了 AETA 系统服务于断裂带测量的需求。对于连续采集数据量过大、数据传输和存储成本高的问题，本文设计实现了两种不同的数据压缩处理算法：（1）对于绝大多数台站，采用了一种基于平滑连续滤波的有损压缩数据处理方法，既保证了数据的连续性，保留了数据中的对地震前兆异常检测和地震勘探中起重要作用的低频信号，又大大降低了数据传输和存储的成本。（2）作为对照和补充，对于少部分台站，采用了一种新的适用于 AETA 电磁和地声数据的无损压缩算法，保留了全部的原始数据信息，并且有效降低了数据传输和存储的成本。该无损压缩算法对 AETA 地声和电磁的原始数据的压缩率分别达到了 0.3 和 0.63 的效果。

2. 将 AETA 原始地声数据可视化，挖掘原始波形中与地震具有较高相关性的异常模式，设计模式识别算法，可以从 AETA 原始地声数据中自动化地识别检测异常波形，并从中提取出新的地声特征。

3. 设计了一种新的转化算法，可以将孤立的地震事件转化为描述地震异常程度高低的标签数值序列，从而实现了地震事件的量化目标，满足了后续将特征与地震事件做相关性分析的需求。

4. 选取了分布于全国各地不同地区的多个台站跨度一年的数据作为实验对象，设计并完成了不同特征与地震事件的相关性分析实验。实验结果表明，本文基于 AETA 原始地声数据提取的地声特征相对于其他电磁和地声特征与地震事件具有更高的相关性。能够更好地服务于地震三要素的预测工作以及模型的建立。

关键词：多分量地震监测系统 AETA，数据压缩，模式识别，相关性分析

AETA data compression processing and analysis on the extracted geoacoustic features

Mengxuan Lv (Microelectronics and Solid-State Electronics)

Directed by Xin'an Wang

ABSTRACT

The prediction of the time, place and magnitude of an earthquake is a worldwide recognized problem. Especially, the short-term and imminent prediction of the three elements of an earthquake can not meet the basic needs of protecting people's lives and property. Under this background, the Key Laboratory of Integrated Microsystems of Shenzhen Graduate School of Peking University developed the multi-component seismic monitoring system AETA, which includes the sensors of geophonic and electromagnetic probes that can collect seismic data, the ground data terminals that can process seismic data, cloud server and data storage system. It can serve for the monitoring of earthquake precursory anomaly, the prediction of three elements of earthquake and the measurement of fault zone. This paper focuses on the data processing and data analysis tasks of AETA multi-component seismic monitoring system, and makes the following work:

1. The improvement scheme of AETA data acquisition process is put forward. The data acquisition method is improved from sampling acquisition to continuous acquisition, which improves the continuity of data. The AETA system not only optimizes the seismic reflection effect of AETA data, but also meets the needs of AETA system for fault zone detection. Two different data compression algorithms are adopted to deal with the problems of large amount of data collected continuously and high cost of data transmission and storage: (1) For the vast majority of stations, a lossy compression data processing method based on smooth continuous filtering is adopted, which not only ensures the continuity of data, retains the low-frequency signal which plays an important role in the detection of seismic precursor anomalies and seismic exploration, but also greatly reduces the cost of data transmission and storage. (2) As a contrast and supplement, for a few stations, a new lossless compression algorithm, which is suitable for AETA electromagnetic and ground sound data, is adopted. It preserves all the original data information and reduces the cost of data transmission and storage. The

compression rates of ground sound and electromagnetic original data are 0.3 and 0.63 respectively.

2. By visualizing the AETA original ground acoustic data, the abnormal patterns with high correlation between the original waveform and the earthquake are mined out. Then, a pattern recognition algorithm is designed to automatically recognize and detect abnormal waveforms from the original ground acoustic data, and extract new ground acoustic features from them.

3. A new transformation algorithm is designed, which can transform isolated seismic events into labeled numerical sequences to describe the degree of seismic anomalies, so as to achieve the quantitative goal of seismic events, and meet the needs of subsequent correlation analysis between features and seismic events.

4. In this paper, the data of a year span of several stations distributed in different regions of the country are selected as the experimental objects, and the correlation analysis experiments of different characteristics and earthquake events are completed. The experimental results show that the features extracted from the original geo acoustic data have a higher correlation with seismic events than other electromagnetic and geo acoustic features. It can better serve the prediction of the three seismic elements and the establishment of the model.

KEY WORDS: AETA, Data compression, Pattern recognition, Correlation analysis

目录

第一章 绪论

1.1 背景及研究意义

地震是一种正常而频发的自然现象。全球每年发生的地震不计其数，其中大多数是微弱而无法感觉到的。但是极少部分的地震由于太过强烈而会给人民的生命和财产安全带来巨大的威胁和破坏。二十世纪全球已经有多达 120 万人因地震而丧生，是造成人员伤亡最严重的自然灾害[1]。我国位于地震带的交汇部位，地震灾害尤其严重。据统计，全球陆地中发生过的地震有三分之一是位于我国境内。为了保护人民的生命和财产安全，预防或减少重大地震灾害带来的严重损失和伤亡，地震短临预测的相关研究具有极大的意义。

地震在孕育和临近发生的时期,会产生各种各样的地声信号。在地震孕育的过程中,随着地应力的逐渐增大，地下的岩石会慢慢地趋于破裂的极限。在这个过程中，岩石层会发生大量的断裂、破碎或者彼此错动,从而会以各种各样形式的信号传播出去，这些传播的信号即为地声信号。同时，地声信号还可以来自于岩石层破裂造成的地下水和泥石流等运动。在这些破裂和错动的同时，还会有地下封闭的气体快速散发而传播出地声信号[2]。地声信号普遍存在于震前、震时和震后，因此前兆地声信号对于地震的预报防灾具有重要的意义。近兆地声信号是临震警报，远兆地声是提供地震预报防灾的可靠信息[3]。

多分量地震监测系统 AETA[4]现阶段已经在川、滇、藏以及河北等地区安装了两百多套地震监测设备，能够连续采集到地下的地声信号。为了满足对地震信号进行实时监测的需求，每个 AETA 台站的数据采集都是连续不间断的。而随着时间的推移，连续采集的地声和电磁数据量也越来越多。为了更好地存储、分析以及利用这些电磁和地声数据，合适高效的数据压缩算法成为一个必要的前提条件。另一方面，由于人类活动以及自然中其他活动的影响，AETA 系统采集到的地声和电磁信号具有信噪比较低、波形复杂多变等不利于进行数据分析的特点，因此设计出有效的模式识别、特征提取算法以便能够从原始数据中提取出与地震具有较高相关性的特征，对于后续基于 AETA 数据的地震预报防灾工作具有极大的意义。

1.2 国内外研究状况

1.2.1 地震预测方法研究现状

地震预测的研究目标不仅要提前预测地震是否发生，还要完成对地震三要素的预

测,包括地震发生的时间、地点和震级。地震的发生机制十分复杂，其发生时间是随机变化的。地震三要素的准确预测是世界公认的难以解决的问题。地震学家根据预报时间跨度的不同，将其分为长期预报、中期预报、短期预报和临震预报四类。其中，临震预报和短期临震预报对防震减灾具有更大的现实意义，是地震预报领域的研究热点。目前，国内外地震预报研究和地震预报的总体理论技术和实践水平还不高，远远无法实现真正的地震预报和地震避灾目标[5]。

总体而言，过去地震预测的研究方向主要有两个：一个是基于历史的地震事件来寻求地震发生的规律进而实现对地震的预测；另一个方向是基于地震的前兆信号来挖掘与地震具有较高相关性的异常特征值进而实现对地震的预测。第一种基于历史的地震事件进行地震预测的方法主要围绕统计学的相关方法，包括简单的概率统计学以及从中衍生出来的机器学习、深度学习等人工智能的方法，都是通过挖掘历史地震事件在空间、时间和能量的三个维度空间的潜在规律和分布特点[6,7]。这一研究方向上，最具有代表性的研究成果是通过概率统计的方法求得古登堡-里克特规则和特征地震分布[8]。而近年来，由于信息技术的进步和地震监测设备的日益精密和发展[9-12]，基于地震前兆信号来实现地震预测的方法越来越成为地震预测研究领域的重心。所谓地震前兆信号，主要包括了电磁信号、地声信号、地下气体液体化学成分的变化、地面隆起或地应变加速等[13-15]。不同的地震前兆信号可以采取不同的数据处理和特征提取方法。1966 年发生邢台地震之后,周恩来总理提出了"地震可以预测"的指导思想,全国进入地震群测群防的阶段。唐山大地震后实现了以专业队伍为依托的地震监测系统，主要包括 6 大学科：流体学科、形变学科、重力学科、应力应变学科、地电学科、地磁学科[16,17]。1987 年尹祥础等学者预测了 1989 年发生在美国加州的 7.1 级地震，其研究理论是通过加卸载响应比理论定量反映震源区介质的变形、损伤过程[18]。对地震前兆信号的研究工作主要是通过运用均值法、中位值法、四分位距法以及滑动窗口等方法来提取并分析原始波形的幅度、信号频率和脉冲能量的大小等相关特征[19-22]。近年来，随着机器学习、深度学习等理论技术的发展，人工智能在地震信号分析和特征提取中的应用也逐渐成为一个热点领域[23-25]。

1.2.2 基于震前地声信号的研究现状

地声是除地震波之外可以直接获取地下信息的重要途经之一[26]。地声信号包括大地震前的震前日常声音以及其他地震前地下传播的高频、不可听的地声信号。地震在孕育阶段，通常会产生地应力场的变化、地壳的缓慢形变、蠕变，由于局部地壳受压而发生形变或微观破裂时，会产生频率范围在几 Hz 至几千 Hz 的地声振动信号；同时，地震 P 波在地震 S 波到达前会引起地下岩层的宏观破裂,产生可听的前兆地声,此外，当地下岩层破裂时，由于高温或其他方式产生的电离气体会在地面附近放电产生地声

信号[27,28]。因此，地声一直是地震预报研究领域中重要的监测和研究的信号。

由于大量的人类活动和自然现象的干扰，地震前兆信号具有复杂性和多变性，对震前地声信号的异常提取工作十分困难。我国关于地震前兆地声信号的研究可以追溯到发生在 1966 年的邢台地震。到 1983 年后又在山东莒县、云南洱源和四川江油等地陆续建立了地声监测台网，使用仪器记录可听地声[29]。1985 年，蒋锦昌等地震学者着力探索地声与地震的关系，他们巧妙地将小白鼠体内 5-HT 代谢过程与综合地声波在传播过程中的衰减特性结合起来，研究得出了地声的信号优势频段[30]。1994 年，郑治真等地震学者首次证实了前兆性地声的存在，他们通过研究分析山东莒县和云南洱源的地声台网观测数据，得出微破裂是地声产生的主要机制的结论[31]。2007 年我国关于地声和地震的大范围研究开启，并建立了基于 GPRS 网络的地声监测系统。近年来，陈维升等地震研究学者基于地声信号的分析和特征提取，挖掘其中与地震具有较高相关性的异常值，取得了对全球范围内较大地震短临预测的较好成果[32,33]。

在 20 世纪初国外已经出现了有关地震地声研究的相关记录。希尔等研究学者长期从事于地声与体波地动关系的研究，1976 年他们研究发现地震 P 波和 SV 波通过震动与空气的耦合是地声信号的成因，推动了地声研究的发展[34]。2007 年，Lin 和 Langston 通过对声音与震源关系的研究，提出了可以依据声音经验性地判定天然地震源的响应特性的结论[35]。三年之后，Lastovicka 等人在地震的震中位置进行了观测实验，采集了次声、地震、磁场和电离层的相关数据，实验结果证明在次声波的激发中起主导作用的是地震运动的垂直分量。2018 年，Shahar 等人开展实验研究分析了意大利中部地震的地声耦合信号，实验结果表明大气、地质环境等因素均会影响地震中地声信号的传播[36]。

1.2.3 时序数据压缩算法的研究现状

数据压缩算法随着信息技术的发展和应用需求的扩大而日益发展起来。按照不同的发展水平划分为四个阶段，分别为发展初期、中期、近期和现状[37-39]。数据压缩算法的第一个发展阶段从 18 世界末到 20 世纪 50 年代，研究认为，最早提出数据压缩思想的人是 Shepards，他在 18 世纪末开展了有关数据压缩的实践活动。直到 1943 年，出现了第一个数据压缩的实例——由莫尔斯发明的莫尔斯电报码[40,41]。数据压缩技术发展的第二个阶段开始于 20 世纪 50 年代。这个阶段数据压缩发展的代表事件是信息熵理论以及香农编码的提出，是由信息论之父克劳德香农在 1948 年提出[42]。三年之后，另一位著名的学者哈夫曼提出了经典的哈夫曼编码[43]。1976 年，J.Rissanen 提出的算术编码进一步推进了数据压缩的研究进展[44]，这种压缩算法使得压缩效果可以逼近信息熵极限。20 世纪 70 年代中期到 90 年代为数据压缩算法发展的第三个阶段。这期间数据压缩逐渐开始成为一门独立的学科，理论和技术也趋于成熟和规范。J.Ziv 和

A .Lempel 两个犹太人在 20 世纪 70 年代中期到 80 年代末期，创新性地抛弃了基于统计思想的压缩算法，发明了基于字典的编码算法,这种更高效更快速的压缩算法被统称为 LZ 系列算法。从 20 世纪 90 年代至今,是数据压缩算法发展的第四个阶段。多学科的发展以及信息技术的广泛应用为压缩技术提出了更高的需求和动力,数据压缩算法进入了一个迅速发展的新时期。2008 年微软推出了 SQL Server2008 社区测试试用版,使得数据压缩技术得到进一步的发展。国内许多高校和研究者也对数据压缩算法做出了很多贡献，其中包括了成为国际标准的音视频编解码标准 AVS[45]。

数据压缩算法包括有损数据压缩和无损数据压缩两大类。数据经过压缩算法压缩处理之后，通过解压算法可以得到与压缩前完全一致的数据的相关算法为无损数据压缩算法。而解压之后并不能得到压缩前完全一致的数据的算法为有损数据压缩算法。相对有损压缩算法,无损压缩算法具有占空间大,压缩比较差的特点,但是可以保证百分之百保留原始的信息,没有数据丢失的问题。无损数据压缩的思想主要有两个方向，一是通过对数据中字符的频率进行统计计算，用编码的方式代替原有的字符，从而达到压缩数据的目的。二是借鉴了查字典的思想，构建 key-value 对，用更短更精简的字符代替原有复杂繁琐的字符实现对数据进行压缩的目的[46]。

其中，第二种构建 key-value 对的基于字典思想的压缩算法由于压缩效果好，效率高，已经成为了主流的无损数据压缩算法。但是在使用过程中由于各种限制依旧会存在各种问题。之后又陆陆续续出现了更多的在基于字典压缩算法基础上的改进压缩算法，主要包括以下三大类：

1) 改进 key-value 对的建立，限制 key-value 对的容量，权衡 key-value 对容量与信息损失之间的关系。

2) 改进 key-value 对的更新，分为抛弃整个 key-value 对或者抛弃 key-value 对中 key 与 value 匹配率较小的节点两种方式。

3) 变换代码的长度，压缩率取决于代码的长度。通过缩短代码的长度，可以有效提高压缩率。

1.3 主要研究内容及方法

目前，AETA 多分量地震监测系统已经在全国各地布设了两百多套地震监测设备，可以不间断地获取电磁和地声的原始数据。根据国内外关于地震前兆数据处理方法的研究现状，为了更好地处理、分析和利用丰富的 AETA 原始数据，并且为后续最终实现地震三要素的预测目标做出更多有价值的贡献，本文主要做了以下两个大方面的研究工作：

1.在数据处理终端层面，介绍了 AETA 系统的组成结构与数据采集流程，通过计算

数据传输和存储的成本，从实际需求出发，总结分析了原有数据采集方案的优势与不足，进而提出了对数据采集流程的改进方案：（1）原先每三分钟取一分钟的抽样采集方案导致了数据的不连续性，大大降低了数据的映震效果，丢失了许多有用的信息，并且不能满足 AETA 用于断裂带检测的需求。放弃抽样采集，新的方案将数据的采集方式改为连续采集，并且提出了两种不同的数据压缩处理方案，减少了数据传输和存储的成本。（2）对于绝大多数台站，采用了一种基于平滑连续滤波的有损压缩数据处理方法，既保证了数据的连续性，保留了数据中的对地震前兆异常检测和地震勘探中起重要作用的低频信号，又大大降低了数据传输和存储的成本。（3）作为对照和补充，对于少部分台站，采用了一种新的适用于 AETA 电磁和地声数据的无损压缩算法，保留了全部的原始数据信息，并且降低了数据传输和存储的成本，对地声和电磁原始数据的压缩率分别达到了 0.3 和 0.63 的效果。

2. 在数据分析层面，通过比较地声和电磁的数据分析成果，分析总结了地声均值特征数据的不足，提出了从地声原始数据出发提取与地震前兆异常相关联的新的特征和分析方法：（1）将地声原始数据可视化，挖掘原始波形中与地震具有较高相关性的异常模式。（2）设计模式识别算法，从地声原始数据中自动化的识别检测异常波形，并从中提取出新的地声特征。（3）设计一种将孤立的地震事件转化为描述地震异常的标签数值序列的转化算法，从而实现了用多种方法计算和比较不同特征与地震事件的相关性，完成了不同特征与地震的相关性分析。

本文主要通过以下方法完成了实验研究内容。

1. 调研以及实地实验

（1）文献调研：查阅分析国内外的相关研究成果文献，较系统、准确、全面地掌握了地震时序数据的压缩处理和特征提取等相关问题和解决方法。

（2）实地实验：通过去四川、西藏等地安装 AETA 地震传感器设备，更加了解了设备安装的周围环境条件及数据的实际物理意义。

2. 多学科交叉融合

通过将物理、信息等多学科理论知识实践相融合，实现了对 AETA 原始数据的合理处理和有效分析、特征提取。

3. 实验分析

通过下载、分析、观察 AETA 采集到的原始数据的特点，调研并比较了多种不同的无损数据压缩算法的优势和不足，设计了适用于 AETA 原始数据的压缩算法。同时，结合实际数据的异常波形设计了模式识别算法，应用于 AETA 系统的原始地声数据中，提取出新的 AETA 地声特征，然后用本文设计的转化算法将地震事件转化为标签数据，进而将提取的特征与转化得到的标签数据进行相关性比对和分析。实验结果表明，本文设计的压缩算法能够有效的对 AETA 原始数据进行压缩，并且从 AETA 原始地声数据中

提取的特征与地震事件有着很好的对应关系，为地震三要素的预测工作做出了很好的贡献。

1.4 论文组织结构

基于本文的研究内容，论文的组织结构如下。

第一章，绪论。首先介绍了地震这一自然灾害给人民的生命和财产安全带来的严重破坏，说明了地震前兆异常检测和地震预测的研究价值。然后介绍了地震预测方法、震前地声信号研究以及可用于地震监测数据的压缩算法的国内外研究现状。最后介绍了本文基于 AETA 系统的原始数据所做的数据压缩处理、地声异常波形分析和特征提取以及特征与地震相关性分析的研究的内容和方法以及论文的组织结构。

第二章，AETA 系统的介绍。首先介绍了 AETA 系统的组成，包括由电磁探头和地声探头组成的地下传感器部分、地面数据处理终端、云服务器和由数据分析人员搭建的数据分析系统。然后详细介绍了 AETA 系统的数据采集流程，包括数据由传感器采集到最终存储到存储器中的详细流程以及不同组成之间的数据交互协议，在充分了解现状之后，提出了数据采集流程的改进方案。

第三章，数据的连续采集和压缩处理。首先结合第二章关于数据采集流程的介绍，通过计算 AETA 原始数据的数据量大小以及低频地震前兆数据的重要性，说明了 AETA 多分量地震监测系统的抽样采集的方案设计的必要性。然后从实际情况和地震监测数据采集的新需求出发，分析了抽样采集所存在的弊端，提出了一种新的不同于抽样采集的连续数据采集的模式。然后进行调研和实验，比对多种不同的数据压缩算法的实际应用效果，设计并应用了适用于 AETA 原始数据的压缩算法，减少了连续采集方案的数据存储和传输的成本，提高了数据分析的效率。最终通过可视化的方法，将连续采集方案的数据波形与地震事件做对比，证明了连续采集方案在映震效果上的优势。

第四章，AETA 原始地声波形异常信号的分析与特征提取。首先介绍了三种基于电磁和地声原始数据计算得到的基础特征数据。分析三种特征数据的优势和不足，发现地声的特征数据丢失了原始波形中的许多与地震相关的信息。通过将 AETA 原始地声数据波形进行可视化操作，分析提取了只存在于地声原始数据中的与地震有较高相关性的异常波形。分析这些异常波形的特点，设计了有效的模式识别算法，进而实现了从原始地声数据中自动化识别异常波形的流程。最后从识别到的异常波形中提取了新的地声特征。设计实验，选取了多个分布在不同地区的 AETA 监测台站的原始地声数据作为实验对象，用本章中设计的模式识别算法和特征提取算法处理这些实验数据，将得到的新的地声特征与台站周围发生的地震事件做对比，发现新的地声特征与地震事件具有很好的对应效果。

第五章，地震事件的标签转化算法和相关性分析。首先调研并分析比较了多种不同的相关性计算算法。然后综合考虑地震的三要素，设计了一种将孤立地震事件转化为异常值标签序列的规范化算法，将多个 AETA 台站周围的一个个孤立的地震事件转化得到了描述地震事件异常值的时序标签数值。然后基于上述研究工作，用多种不同的相关性计算方法设计实验并计算电磁、地声特征和本文提取的新特征与地震事件的标签序列之间的相关系数。对实验结果进行了总结和分析。

第六章，总结与展望。首先总结了本文在 AETA 系统的数据处理和数据分析系统所做的工作和创新。然后结合研究生期间参与的项目，对后续的研究工作进行了规划和展望。

第二章　AETA 系统的介绍

第一章从本文工作的研究意义和背景出发，详细介绍了地震预测、震前地声信号以及可用于 AETA 原始数据压缩的算法的国内外研究现状。本章将详细介绍北京大学多分量地震监测系统 AETA 的组成结构以及数据采集的详细流程。在充分了解 AETA 系统的基础上，结合地震监测的数据连续性需求、断裂带测量的需求以及数据传输和存储的成本需求，提出了新的目标：更充分高效地采集地震数据，减少数据存储传输的成本，提取更有效的、与地震事件具有更高相关性的特征。

2.1　AETA 系统的组成介绍

AETA 多分量地震监测系统是由北京大学深圳研究生院集成微系统实验室研发并布设。作为一套可以大区域、高密度布设的地震监测数据采集的软硬件综合体系[4]，该系统可以用来观测地震前兆异常以及震时相关数据情况。其组成部分如图 2.1 所示。

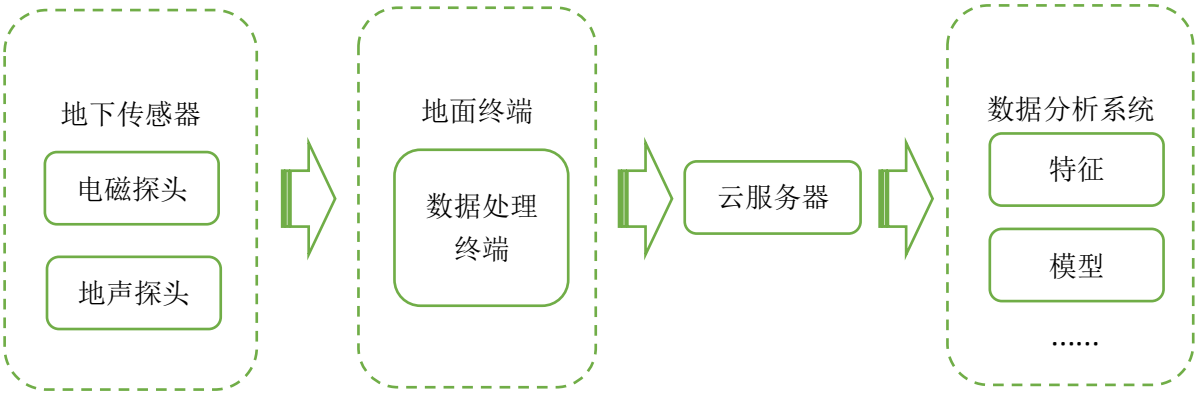

图 2.1 AETA 系统示意图

2.1.1 地下传感器

AETA 地下传感器包括电磁探头和地声探头两部分。电磁探头是一种电磁扰动传感器，磁场在垂直方向的变化可以使其得到感应电动势。经过放大、滤波和数模转换等处理可以上传到数据处理终端得到电磁信号[47-49]。电磁频率的响应范围为 0.1 赫兹到 10k 赫兹，可以监测到的电磁数据的范围是 0.1nT 到 1000nT，灵敏度高于 20mV/nT@0.1 赫兹~10k 赫兹，低频采样的频率为 500 赫兹，高频采样的频率为 30k 赫兹。地声传感器是一种压电薄膜传感器[50,51]，通过感应地声信号产生形变进而感应出电荷。经过电荷放大器、滤波器、放大电路和数模转换等处理得到地声信号。其频率响应为 0.1 赫兹到 10k 赫兹，电压分辨率为 19.073uV,幅值一致性误差小于 5.5%，高低频的采样频率与电

磁探头一致，分别为 500 赫兹和 30k 赫兹。电磁和地声探头的外形如图 2.2 所示。

图 2.2 电磁和地声探头示意图

2.1.2 地面终端

AETA 地面终端为数据处理终端，可以处理由地下传感器采集到的各种电磁和地声的数据。数据处理终端主要完成对两个传感探头数据的实时接收、处理和发送，同时还要接收并处理来自服务器端的应用程序的指令，并且实现对探头的控制管理。

图 2.3 地面终端示意图

数据处理终端在硬件部分采用的机箱设计是标准的 2U，可以存在机柜中。液晶、存储器和单片机三部分组成的 GPU45A 串口屏位于终端面板左边，复位键和电源键位于右边。电磁、地声传感器的网线插口，网络通信接口，两个 USB 插口，调试串口和数据串口从左到右依次排列在终端的背部。机箱内部包含有 M3352 工控板、交换机和电源模块电路。

软件部分，数据处理终端与电磁和地声传感器之间采用的是 C/S 模式设计。二者通过 TCP/IP 传输层协议进行通信。可以支持多探头并发请求，能够根据需求设置多个缓冲区来存储多个不同传感器上传的待解包的数据。同时，数据处理终端还可以向地下传感器下发命令、配置参数等。在数据采集流程中，数据处理终端还要负责对数据进行压缩处理操作，并将处理好的数据发送到上层服务器的应用程序中。在与上层服务

器的应用程序进行通信时，数据处理终端可以主动发送数据，并且实现了多个功能，其中包括：命令查询与请求、配置、状态更新、日志上传、实时发送警报以及同步时间等。通过 HTTP 协议，数据处理终端还可以向上层的服务器应用程序发起多种类型的数据请求。

2.1.3 云服务器

云服务器主要负责与地面终端进行通信，并提供任务服务、数据分拣服务和监控服务。AETA 云服务器采取了负载均衡集群技术部署，可以将访问流量分散并发，从而可以接入更多的监测台站，也可以避免因为单点的故障造成的数据缺失。

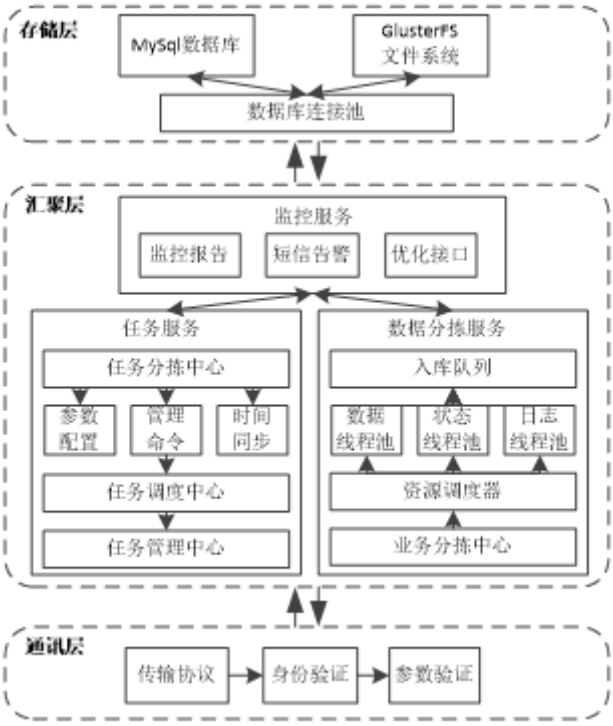

图 2.4 AETA 服务器系统框图

云服务器的整体架构如图 2.4 所示，主要由通讯层、汇聚层和存储层三部分组成：

（1）通讯层负责通信协议的解析，可以将来自终端的请求接入到云服务器中，保障了终端与服务器之间的数据通信。

（2）汇聚层负责三类服务，包括任务服务、数据分拣服务和监控服务。任务服务主要是将操作命令、参数配置和时间同步等任务派发给终端，从而实现了远程管理终端；数据分拣服务是接收和处理电磁、地声数据和设备状态、日志、告警等信息；监控服务实时监控所有台站设备和 AETA 服务器。

（3）存储层负责持久化存储来自汇聚层的数据。

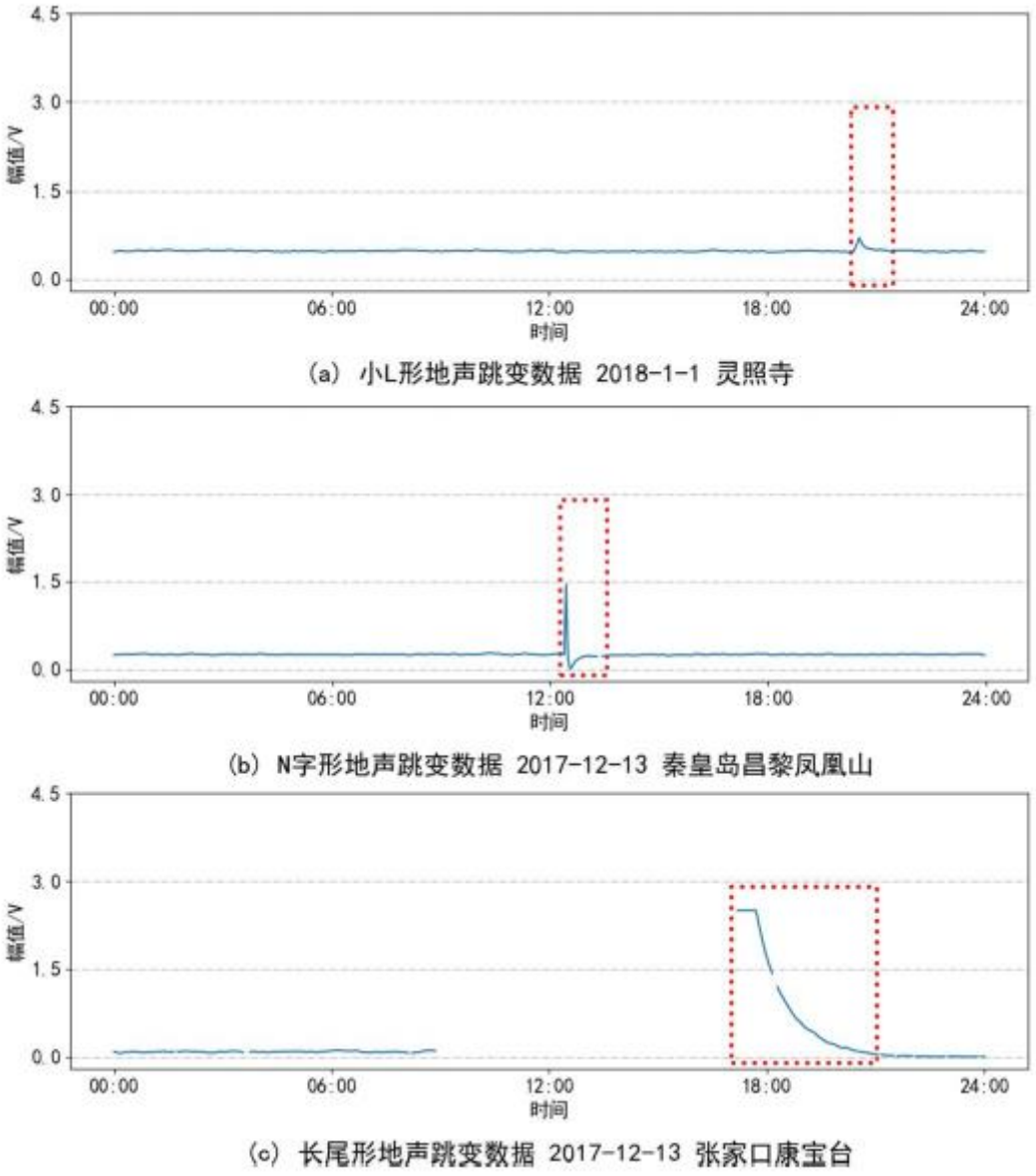

图 2.5 断电导致的地声数据跳变示意图

2.1.4 数据分析系统

数据分析系统主要是实验室的数据分析同学在老师的指导下，从特征、规则、模型等角度出发，以电磁和地声的均值、振铃计数和峰值频率三种特征数据以及电磁和地声的原始数据作为研究对象，以地震三要素的预测为最终目标开展的数据处理和分析的相关工作。

数据分析系统主要围绕数据的预处理、特征的提取、模型的建立三方面展开工作：

1.数据的预处理，包括了断电数据的处理和缺失值的处理。

（1）为了更好的监测并采集地震数据，AETA 多分量地震监测系统广泛布设于各种区域，在一些偏远山区，部分台站可能会出现断线重启的情况。其中一些早期设备存在

一个问题，即在设备经过断电而重启后，地声数据中会出现一种异常的跳变现象，数据波形如图 2.5 所示。这种早期设备由于断电而产生的异常跳变信号与一些其他地震等自然现象造成的强烈地声信号相类似，因此如果不采取一些恰当的数据处理方法对其进行合理的处理，会极大地干扰并影响到后续的数据分析、特征提取和模型建立等工作。AETA 数据分析系统根据这些跳变信号的特征，实现了一个自动识别断电造成的跳变数据的模式识别算法，通过自动识别这些断电数据，对其进行了删除操作。

（2）缺失值的处理是通过一些算法和规则来减少缺失的数据对后续分析工作的影响。数据的缺失问题如果不能够有效的处理，会造成诸多问题，例如使数据的统计特征和分布特征对数据的估计产生偏差。常用的处理缺失值的方法主要有：基于已有数据统计的均值插补、多重插补、极大似然估计；基于模型的建模预测、高维映射等[52-54]。

AETA 的数据缺失的因素有很多，主要包括了部分偏远地区的断电和断网等问题以及偶尔情况下的系统故障等。基于快速有效、减少运算成本的原则，一般采用的是一些均值插补和线性插补等简单而快速的缺失值处理方法。

2. 特征的提取，包括从地声、电磁的原始数据和三种基础特征数据中通过规则、无监督模型等方法提取出与地震具有较高相关性的高阶特征。现阶段数据分析系统的特征提取工作已经取得了一些与地震具有较高相关性的特征，包括以下几大类：

1）基于规则的特征：电磁均值的 SRSS 波形、地声均值的区域性显著异常特征。

2）基于无监督算法：滑动主成分分析法[55]提取 SRSS 波形异常值、局部互相关跟踪法[56,57]提取电磁均值异常值、滑动四分位距法[58-61]提取电磁均值异常值以及分型维数算法[62-64]等。

这些特征的提取将数据量巨大、抽象的原始数据和基础特征数据，转化成了与地震具有较高相关性的高阶特征，从组成复杂、信噪比低的电磁和地声信号中，挖掘出了与地震具有较高相关性的信息，为后续模型的建立、地震三要素的预测提供了很好的特征基础。

3. 模型的建立。AETA 数据分析系统关于地震的短临预测模型，要严格以地震三要素的预测作为目标。模型与数据预处理、特征提取的工作成果紧密结合，并提出了分层数据处理的方法，将数据预处理、特征提取、异常检测等工作进行分层处理完成。而地震三要素的预测采取了独立预测的方法：

（1）通过构建风险及位置特征集将地震风险与地理位置相关量，通过聚类等算法预测震中。

（2）锁定地点和时间，将震级进行分箱处理转化为多分类模型进行预测震级。

（3）固定地点，筛选震级，将时间进行分箱处理转化为多分类模型预测地震发生的时间范围。

2.2 AETA 系统的数据采集流程

AETA 的数据是由电磁和地声传感器采集，然后经过数据处理终端经过抽样、压缩、滤波等处理之后，发送到云服务器进行多任务处理，并最终存储到数据中心当中。具体流程如图 2.6 所示。

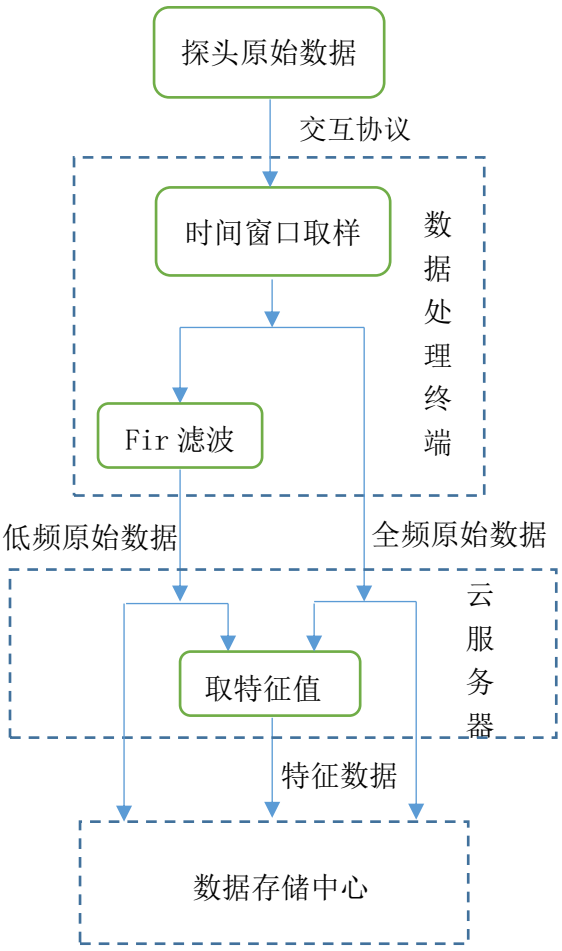

图 2.6 AETA 数据采集流程示意图

在地下传感器和数据处理终端之间通过 TCP/IP 协议进行通信，数据处理终端与服务器的应用程序之间通过 HTTP 协议通信。由于数据来源、类型、时间以及长度等信息都需要数据处理终端和服务器进行识别，因此需要设计出一种合适的数据传输协议，以保证地下传感器与数据处理终端、数据处理终端与服务器的应用程序之间的数据传输能够有序完成。

地下传感器与数据处理终端在成功建立连接之后，便开始向数据处理终端发送采集到的电磁和地声数据。由于地下传感器采集的数据会以相当快的速度向数据处理终端发送，因此在下一条的数据发送之前并不会等待数据处理终端的回复消息。除此之外，地下传感器还会定时将包括自身的运行状态，传感器软件的版本号、发送间隔等信

息发送给数据处理终端。与此同时，数据处理终端也会将一些复位、重启、更新等控制命令发送给地下传感器。因此，地下传感器与数据处理终端之间的信息交互是双向。二者之间的交互协议如图 2.7 所示。

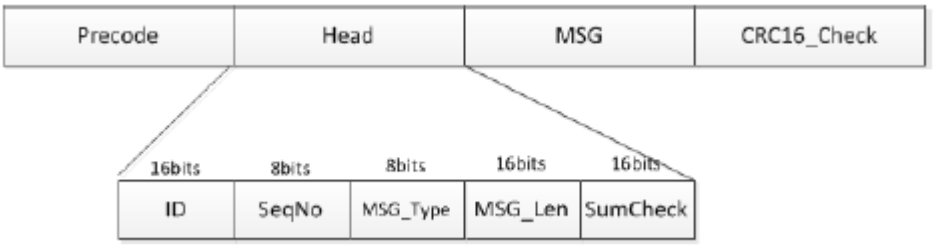

图 2.7 数据终端与传感器之间的数据协议

主要由以下几部分组成：

（1）Precode 前导码：数据包最开始的部分是 Precode 前导码，其字节偏移为 0，因为前导码需要与数据包其他的组成部分相异，因此暂定为 0xFFFF，占据两个字节大小的空间。

（2）Head 头：由 ID,SeqNo,MSG_Type,MSG_Len 和 SumCheck 五部分组成。其中 ID 表示探头的设备编号;SeqNo 是数据包的序列号，随每一个数据包的生成加 1;MSG_Type 用于区分消息的类型，0x01 表示数据，0x02 表示命令；MSG_Len 代表 MSG 的长度，即包含有多少个 16bits 的数据；SumCheck 为包头和校验码。

（3）MSG 消息：MSG 消息有数据和命令两种类别，格式如图 2.8 和图 2.9 所示：

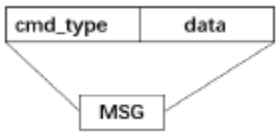

图 2.8 命令的消息格式

图 2.9 数据的消息格式

地下传感器上传到数据处理终端的数据有多种数据类型，包括电磁探头数据、地声探头数据、温度数据、探头状态和探头日志等，图 2.10 详细说明了不同的 data_type 所代表的含义。

表 2.1 data_type 数据含义

data_type	说明	数据说明
0x0001	电磁数据	18 位数据低两位不要，只传高 16 位，不间隔连续传输，一次传输 1024 个点，数据为有符号数据
0x0003	地声数据	18 位数据低两位不要，只传高 16 位，一次传输 1024 个点，数据位有符号数据
0x0007	温度数据	16 位，温度数据 1min 一个值，一次传输 1 个温度值，换算时除以 100 则可得到温度值
0x0009	探头状态	ID：16 位；探头类型：8 位，1 为地声，2 位电磁，其他保留；软件版本：24 位，A.B.C；参数同步间隔周期：16bits；零漂补偿值：16bit；放大倍数：16bit。顺序为 ID+类型+软件版本+NULL_复位+参数同步间隔周期+零漂补偿值+放大倍数。
0x000a	探头日志	0x0001：探头重启；0x0004 表示 crc16 校验出错；0x0005 表示和校验出错；0x0006 表示前导码错误

Cmd_type 为 0x0002 时代表 MSG 消息是命令的形式，地下传感器与数据处理终端之间命令形式的 MSG 消息交互是单向的，只能由数据处理终端发送给地下传感器。当 data 为 16 进制的 0 时，地下传感器禁止复位；为 1 时表示命令传感器进行复位操作；为 2 时表示命令传感器停止数据的发送；为 3 时表示命令传感器发送数据。同时，MSG 消息的功能还可以继续拓展。

在数据处理终端中，传感器上传的数据量庞大的原始数据会进行间隔抽样的处理，抽样的方法是通过窗口取样，然后用一个临时存储空间进行暂时存储。临时存储空间内数据会做两个不同的处理流程：其中一路会做滤波处理，将原始数据滤波得到频率为 500 赫兹的抽样数据；另一路会直接通过 HTTP 协议发送到应用服务器中。

数据处理终端在与应用服务器通信的过程中，采用的协议是 HTTP 协议。二者之间数据和参数的传输方式是 POST 终端向服务器发送 http 请求。而日志和原始数据这种大量的数据会以文件的形式上传，上传的信息包括：时间戳、终端号、数据类型和协议版本号等构成了文件名。以 1501652352_38_1_1 为例，1501652352 代表时间戳，38 为终端号，数据类型和协议版本号均为 1。

应用服务器同样会对数据处理终端上传的数据做两种处理流程。其中一路会直接存储起来；另一路会做进行均值、振铃计数和峰值频率的特征值计算和提取，然后将提取到的三种特征值存入到数据存储层中。

2.3 本章小结

本章第一节详细介绍了 AETA 多分量地震监测系统的组成，包括负责采集数据的地下电磁和地声传感器、负责进行数据处理的地面数据处理终端、云服务器以及最终的数据分析系统。第二节详细介绍了 AETA 的整体数据采集流程。包括地下传感器与数据处理终端、数据处理终端与云服务器之间的交互协议，原始数据的存储和三种基础特征数据的计算、提取和存储流程。第三章将围绕 AETA 数据采集流程中涉及到的一些可优化的点，详细介绍本文关于数据采集流程的一些改进和创新的工作。

第三章　数据的连续采集与压缩处理

上一章介绍了在地下传感器采集到数据之后，由地面数据处理终端对数据进行采样和处理。在系统设计之初，根据现场实验所获得的数据特征以及数据量等相关信息，在软件设计中采用了抽样采集的方式。本章将通过分析原始数据的数据量和数据特点，提出了一种改进的连续采集的数据采集方案，并根据连续采集方案中遇到的数据量等问题，设计出一种合适的无损数据压缩算法，最终在满足了数据的连续采集需求的同时，降低了数据传输和存储的成本。

3.1　AETA 原始数据量分析

AETA 多分量地震监测系统的原始数据由电磁和地声原始数据组成。电磁和地声传感器按照 ADC 的采样频率可以分为低频和高频两种版本，表 3.1 分别计算和统计了不同类型探头的数据量。

表 3.1 原始数据的数据量统计表

探头类型	采样率 (kHz)	数据量 (MB/分)	数据量 (GB/时)	数据量 (GB/天)	数据量 (GB/月)	数据量 (GB/年)
高频电磁探头	30	3.43	0.201	4.82	144.6	1735.2
高频地声探头	150	17.15	1.005	24.1	723	8676
低频电磁探头	10	1.14	0.067	1.61	48.2	578.4
低频地声探头	30	3.43	0.201	4.82	144.6	1735.2

从表中可以看出，那么一个 AETA 台站每年会产生大约 10TB 的原始数据，如果不对这些庞大的原始数据进行合理的压缩处理，会消耗大量的宽带流量和存储资源。并且 AETA 多分量地震监测系统在全国各地已经布设了两百多套传感器设备并将继续进行大区域高密度布设，因此，合理有效降低数据传输和存储的成本是必不可少的。

为了有效的减少数据传输和存储的成本，AETA 系统在设计之初采用了抽样采集的数据采集方案，可以极大的减少数据的量，满足了数据传输、存储低成本的要求，有利于大区域高密度布设 AETA 地震监测台站。但是抽样采集方案丢失了很多有用的信息，破坏了 AETA 数据的连续性。

3.2 原始数据的连续采集方案设计与分析

3.2.1 连续采集方案的需求分析

抽样采集方案虽然有效降低了数据传输和存储的成本，但是在运行的过程中发现了以下几个存在的问题：

（1）每三分钟抽样窗口丢失两分钟的数据，丢失的数据可能包含有价值的地震前兆信息。

（2）在台站周围发生地震之后，由于抽样采集数据的不连续性，使得每三分钟只有一分钟的数据被采集并经过滤波上传到云服务器中，因此大多数台站在周围发生地震时并不一定会被采集到实时的数据，因此多数台站的数据没有较好的映震效果。

（3）AETA 系统在进行地震前兆预测的同时，后续还要再断裂带的测量工作中发挥重要作用，数据的连续性要求提高。

因此，有必要将数据的采集方式从抽样采集改为连续采集。连续采集虽然可以满足数据连续性的要求，但是会陡然增加数据的量，极大的增加了数据传输和数据存储的成本。正如 3.1 节中计算统计所得，每个台站每年会产生多达 10TB 的数据量。设计并应用一个合理有效的数据压缩算法来降低数据的量是极为必要的。同时，为了满足不同的数据传输、存储和分析的需求，连续采集方案也对不同的 AETA 台站采用了不同的数据压缩处理方法。对于绝大多数台站，采用了一种高效的连续滤波有损压缩处理方式；作为对照和补充，对于少部分台站，设计并应用了一种无损的全频数据压缩的连续采集方案。

3.2.2 连续采集方案设计

基于 AETA 的数据分析研究成果表明，地震前兆异常主要集中于低频范围中，特别是在 200 赫兹以下的低频数据。并且低频的地声数据具有传输范围远、穿透力更强等特点，在地震前兆异常监测和断裂带检测中可以发挥更加重要的作用[65,66]。因此，在数据处理终端中，放弃原先的抽样采集数据的方案，对接收到的探头数据进行连续滤波，得到数据量较小的低频连续数据，在满足数据连续的需求的同时，也规避了传输带宽和存储量的问题 。其中，滤波方案的具体步骤如下所示：

（1）滤波器采用的是 64 阶滤波器，每 64 个点计算出一个低频数据。

（2）滤波线程每次取 1s 钟的数据作为滤波算法的输入进行滤波。

（3）为凑够 64 个点而无法进行滤波的数据，将于下一秒钟的数据拼接输入。

这种平滑连续滤波的方案，可以保证滤波之后数据的连续性。

低频滤波连续采集方案虽然保证了数据的连续性，但是仍然会丢失一部分数据和信息，因此为了确保重要前兆异常信号的完整性，在一些重点的 AETA 台站还部署了全

频连续采集版本。这种采集方案主要通过一种无损压缩的算法，数据处理模块即为压缩线程，压缩线程将地下传感器上传的 1 分钟的原始数据进行压缩处理之后，放入到发送缓冲中，由发送线程将压缩后的数据直接发送到服务器当中。无损压缩算法的实现将在下一节中详细描述。

3.3 原始数据的无损压缩处理

3.3.1 多种无损压缩算法的分析与比较

无损压缩算法又称为信息保持编码、无失真编码、熵编码等等，是可以根据一定的规则对大量的数据进行编码压缩，在压缩之中不会出现精度的损失，并且被压缩的数据可以通过解压算法完整还原[67]。无损压缩算法又可以根据编码原理分为两大类：

1. 基于统计压缩算法：这种压缩算法出现较早，主要是通过统计字符的频率来进行字符编码。主要包括以下几类：

（1）香浓-凡诺算法[68]。这种算法是通过二叉树来标注每个码元的出现频率，把短编码分配给高频率的符号来实现数据的压缩，具体逻辑如图 3.1 所示。算法首先会根据原始数据集来统计出每个字符出现的频率；然后根据字符频率的大小对字符进行排序，频率高的字符位于左边，频率低的字符列于右边，如表 3.2；之后再把排序好的字符列表分为两部分，保证左边部分的总频率数接近于右边部分；用数字“0”分配给左半部分的二进制数，数字“1”分配给右半部分的二进制数；然后将最左和最右两部分分别重复上述操作，每重复一次都会添加一维编码，直到每部分只含有一个字符为止。

表 3.2 香浓-凡诺编码字符频次与频率统计示意表

符号	A	B	C	D	E
频次	15	7	6	6	5
频率	0.3846	0.1795	0.1538	0.1538	0.1282

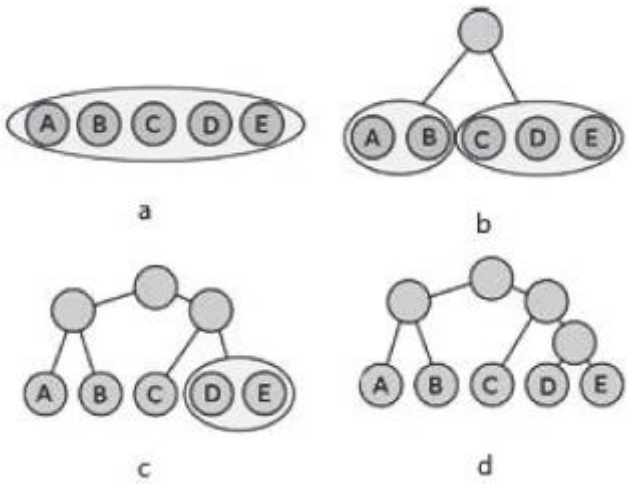

图 3.1 香浓-凡诺编码流程示意图

由流程图可以得到编码如表 3.3 所示：

表 3.3 香浓-凡诺编码

符号	A	B	C	D	E
编码	00	01	10	110	111

因此可以计算这段数据的码字平均长度为：

$$\frac{2\ \text{Bit}\cdot(15+7+6)+3\ \text{Bit}\cdot(6+5)}{39\ \text{Symbol}}=2.28\frac{\text{Bit}}{\text{Symbol}} \tag{3.1}$$

（2）哈夫曼编码[69]。其核心思想是用最短的编码为出现频率最大的字符进行编码，从而达到每个符号平均比特数尽量小的目的。算法的步骤首先是将原始数据中不同字符按照频率大小进行排列；然后将频率最小的两个字符所对应的频率相加，作为新的频率；然后重新排序，重复上述操作，直至最后两个频率之和为 1；从下到上构造编码二叉树，根据生成的二叉树的结构得到符号对应的编码。

依旧按照表 3.2 的实例可以得出哈夫曼编码如图 3.2 所示：

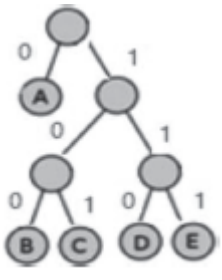

图 3.2 哈夫曼编码流程示意图

由流程图可以得到编码如表 3.4 所示：

表 3.4 哈夫曼编码

符号	A	B	C	D	E
编码	0	100	101	110	111

因此可以计算这段数据的码字平均长度为：

$$\frac{1\ \text{Bit} \cdot 15 + 3\ \text{Bit} \cdot (7+6+6+5)}{39\ \text{Symbol}} = 2.23\ \text{Bit/Symbol} \qquad (3.2)$$

（3）动态哈夫曼编码。这种算法是对哈夫曼编码的改进。通过取消统计过程，压缩过程与动态调整哈夫曼树同时进行，节省了统计过程的计算时间和资源的消耗。从而使得运行时间与输入串长度成线性关系，空间需求为常数。

（4）算术编码。这种编码方式是用[0,1)区间来代替将被编码的数据序列，间隔的位置由输入数据的频率所决定。当需要表示的信息很长时，需要的二进制位数就会增多，区间间隔也就越小。编码时首先会根据字符在原始数据中出现的频率，将[0,1)区间分成多个区间段；将[0,1)区间设置为初始区间；根据待处理的数据信号，一个个读入信号源，每次读入一个，就将该信号源在[0,1)区间上等比例缩小到最新的区间中；重复上述步骤直到读完全部的数据信号。

基于统计的压缩算法各有所长。表 3.5 比较了上述四种压缩算法的优缺点。

表 3.5 统计编码优缺点比较

算法名称	优点	缺点	适用范围
游程长度编码	实现简单；压缩和解压速度快	适应性差；平均压缩率低	复杂度低的原始点阵图像
哈夫曼编码	简单实用；编码译码唯一	受被压缩文件大小的影响大；速度慢	GZIP,JPEG 数据
动态哈夫曼	提高了速度	只能去除概率分布不均匀的冗余	字符出现概率不均匀和大数据的压缩
算术编码	压缩率最高	运算复杂速度慢	信源概率比较接近

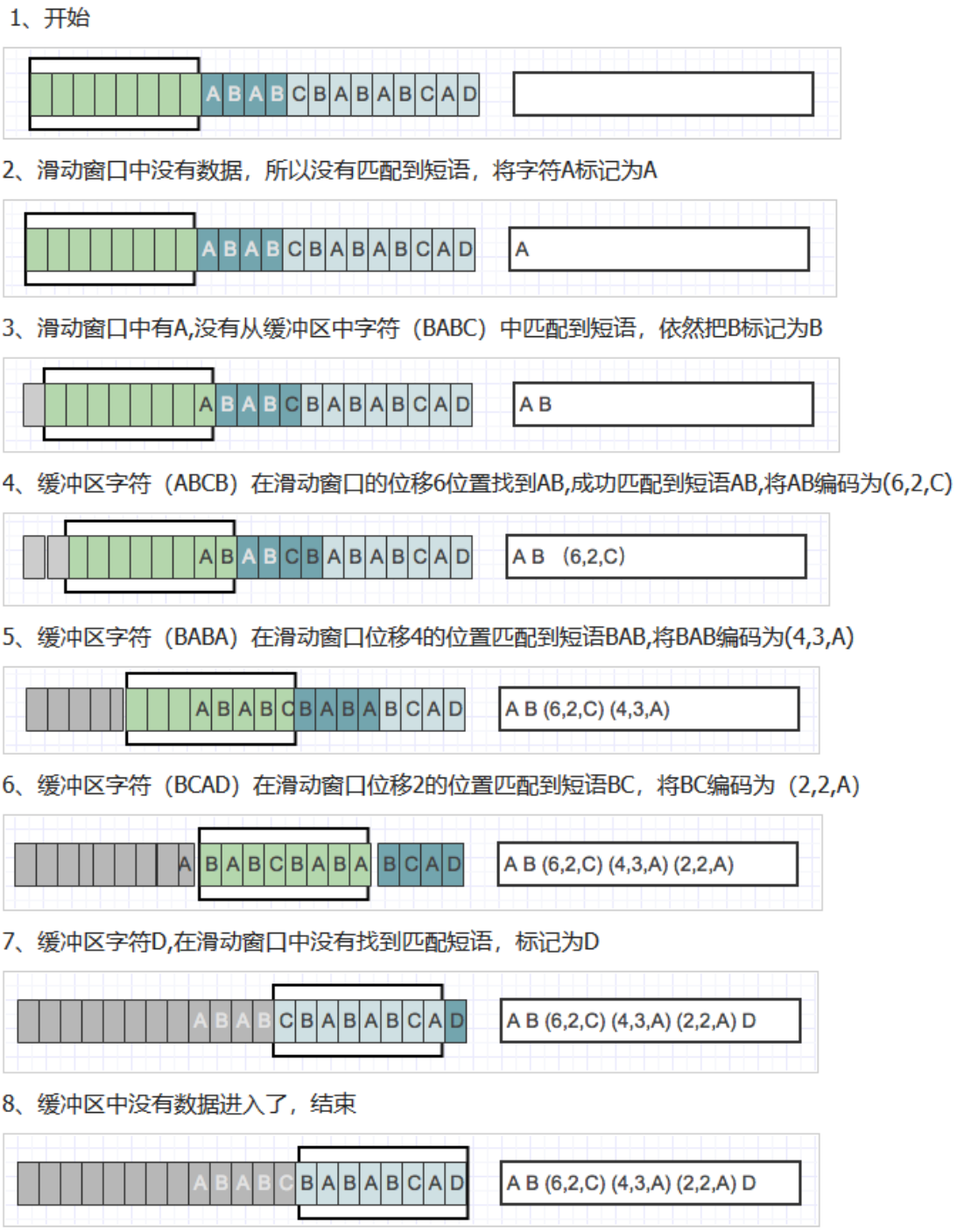

图 3.3 LZ77 算法流程示意图

2. 基于字典压缩算法：这种压缩算法是借鉴了查字典的思想。首先会建立 key-value 对，key-value 对中的 value 由较长的字符串或者经常出现的字符组合构成，然后 key 由较短的符号来表示。主要包括以下几类：

（1）LZ77 算法[70,71]。这种算法是通过滑动窗口的方式实现的，使用一个前向缓冲区和滑动窗口，具体流程如图 3.3 所示。

（2）LZ78 算法[72,73]是在 LZ77 算法上的一种改进算法。它通过构建一种树形结构的 key-value 对代替滑动窗口来保存短语。在开始时，前缀 A 和 key-value 对均为空。读取字符 B，判断 key-value 对当中是否存在 A 与 B 相连的字符串。如果是，将 B 接入到 A 当中；如果不存在，输出 key-value 对中 A 所对应的码字 B，然后把字符串 A+B 添加到 key-value 对中，同时将 A 置空。重复上述操作，直至字符流冲为空；输出相应于当前前缀 A 的码字，压缩流程结束。

（3）LZSS[74]算法是一种改进 LZ77 性能的使用算法，采用的是二叉搜索树，可以提高压缩速度并且在解码过程中不需要生成和维护二叉搜索树。

（4）LZW[73,75,76]算法是一种对 LZ78 算法进行改进的算法，主要适用于 GIF 格式的图像压缩。在对数据流进行压缩过程中，会重复出现很多词汇和短语，因此重复的序列可以用一个短的编码来代替。LZW 流程如图 3.4 所示。

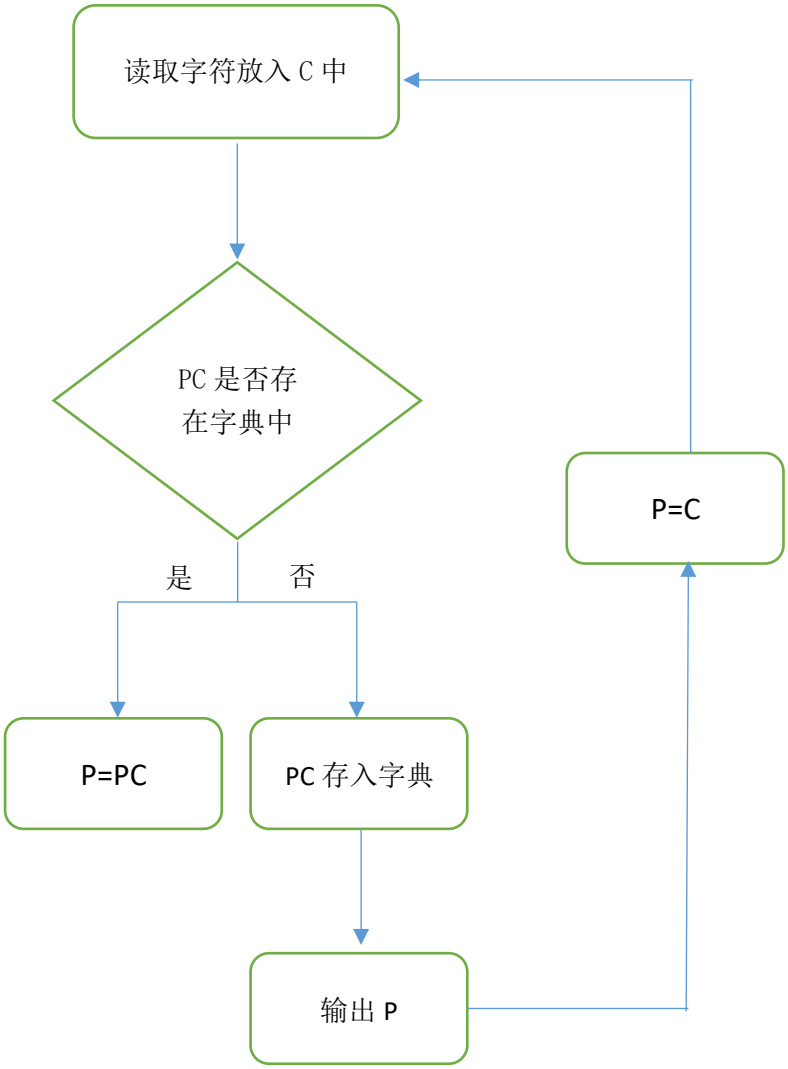

图 3.4 LZW 算法流程示意图

基于字典的压缩算法目前成为主流的无损压缩算法，但在实用过程中，不同的字典压缩算法依旧各有所长，总结如表 3.6。

表 3.6 统计编码优缺点比较

算法名称	优点	缺点	适用范围
LZ77	基础算法，压缩效果好，速度快	压缩率较低，空间局限性	单片机
LZ78	效果好，速度快，树形结构确保字典包含全部内容	编译复杂，实现困难	具有一定区间重复性的数据文件
LZSS	提高了解压速度，编译简单	每次压缩需要搜索开头，不利于较长原文的编码；存在冗余	较短原文
LZW	效果好，速度快，编译简单	存储空间与输入长度成正比，不利于压缩性能	GIF 图像；txt 文本

3.3.2 AETA 原始数据的压缩处理

通过调研和分析上述几种基础的压缩算法可以看出，压缩的思想在于去除空间冗余，用短编码代替高频率字符，用少量编码代替长串重复字符，从而降低原始数据文件的大小。另一方面，AETA 的原始数据具有以下显著的特点：

（1）均为数值类字符。

（2）绝大多数数据的高位均为 0。

（3）重复子串较少。

（4）采样频率高，相邻数值差值小。

结合这三种特点，什么样的算法可以更好的更有效的进行压缩处理呢？本文提出了一种基于数据位变换的无损压缩算法，通过相邻求差法、正负转换以及位变换，可以将重复子串较少的原始数据转化成了重复子串出现较多、且长度较长的原始数据。具体步骤可以描述为：

（1）相邻求差值：由于原始数据相邻两个数值的差值极小，因此可以保留原始数据中第一个数，将后面的数值依次与前一个数值进行求差，用差值序列代替原始数值序列。进而得到了一个绝对值极小的数值序列。

（2）正负转换：为了保证最高位，即符号位的数值均为零，对差值序列进行正负

转换，得到无符号整数序列。为了保证正负信息的不丢失，方便解压，正负转换可以通过奇偶与正负相映射的方式实现。即负数 m 转换为 2m-1，正数 n 转换为 2n。转换公式如下：

$$f(x)=\begin{cases}2x, x\geq 0\\ 2|x|-1, x<0\end{cases} \tag{3.3}$$

其中 f(x)表示原始差值 x 经过正负转换之后得到的无符号整数。

（3）位重组：将无符号整数序列中每一个数值按照位进行拆分，然后进行重组。即依次获取每个数据的最高位，然后再获取次高位，重复操作直至最后一个数据的最低位。将位重组得到的新的数据重新组合成新的同样位数的数值序列，因为无符号整数序列中绝大多数数据的高位均为 0，因此新得到的数值序列存在大量的重复子串，也就是 0 子串，极大的方便了数据的压缩处理。

（4）数据编码：经过前面三步的处理，已经得到了极其容易压缩的数值序列，编码可以采用最简单的方式。顺序遍历位重组之后的数值序列，统计连续出现多少次 0 值，用重复个数和两个标志位来代替重复的一串 0，对于非 0 值不做处理。

通过上述四步，可以将 AETA 原始数据进行快速有效的压缩处理，流程如图 3.5 所示。

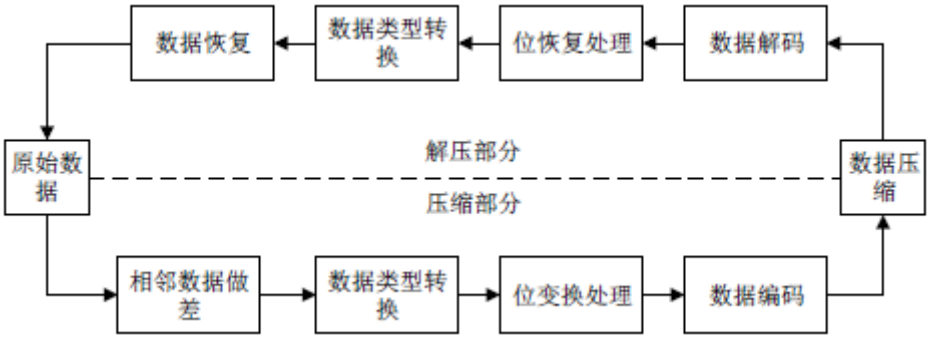

图 3.5 AETA 数据无损压缩和解压流程示意图

解压缩的过程即是将压缩之后的数据，根据压缩的步骤逆向反推得到原始数据。首先根据数据编码的规则，找到编码的标志位和当前子串 0 的个数，恢复成重复的 0 子串，而没有编码标志位的子串则不做处理、重复拷贝。解码之后，再根据正负转换的规则将无符号数值序列恢复成有符号数值序列，最终逆向操作差值即可解压得到原始数值序列。

3.3.3 AETA 原始数据的压缩效果

将上一节中描述的无损压缩算法应用到 AETA 多分量地震监测系统中的电磁和地声

原始数据中，随机选取了 12 个台站的电磁和地声数据作为实验对象，经过实验和计算得到如下实验结果：

表 3.7 电磁数据压缩效果表

台站名称	终端号	探头号	压缩倍率	压缩比
泸定气象局	22	24	1.36	0.73
康定姑咱山洞	26	56	1.78	0.56
九龙县乃渠乡	130	112	1.61	0.62
石棉挖角乡	33	31	1.24	0.81
石棉山洞	35	30	1.34	0.74
宝兴县灵关中学	93	106	1.45	0.69
名山区安吉村	124	133	1.23	0.81
西昌气象局	32	19	1.45	0.69
冕宁防震减灾局	38	22	1.21	0.83
德昌防震减灾局	41	99	3.59	0.28
盐源县盐塘乡	78	123	1.68	0.59
雷波县地震台	84	118	2.03	0.49

表 3.8 地声数据压缩效果表

台站名称	终端号	探头号	压缩倍率	压缩比
泸定气象局	22	30025	3.28	0.30
康定姑咱山洞	26	30017	3.52	0.30
九龙县乃渠乡	130	30135	3.62	0.28
石棉挖角乡	33	30023	3.28	0.30
石棉山洞	35	30022	2.39	0.42
宝兴县灵关中学	93	30079	3.38	0.30
名山区安吉村	124	30141	3.59	0.29
西昌气象局	32	30019	3.29	0.30
冕宁防震减灾局	38	30021	3.31	0.30
德昌防震减灾局	41	30133	3.47	0.29
盐源县盐塘乡	78	30138	3.50	0.29
雷波县地震台	84	30123	3.66	0.29

根据这 12 个台站的电磁和地声数据实验计算可得：电磁数据的平均压缩率可达 0.63，地声数据的平均压缩率为 0.3。每个台站一个月的数据量，可以由原先 197.7GB，压缩到 75.5GB，有效降低了数据传输和存储的成本。

3.4 连续采集方案的映震效果

将数据的采集方式从抽样采集改为连续采集之后，因为不再存在丢失数据信息的情况，数据的连续性得到极大的加强，因此 AETA 数据对地震事件的映震效果也得到了明显的提高。如图 3.6 所示，经过简单的算法提取，可以看到 AETA 台站对周围的地震事件能够有明显的对应的高峰值显示。

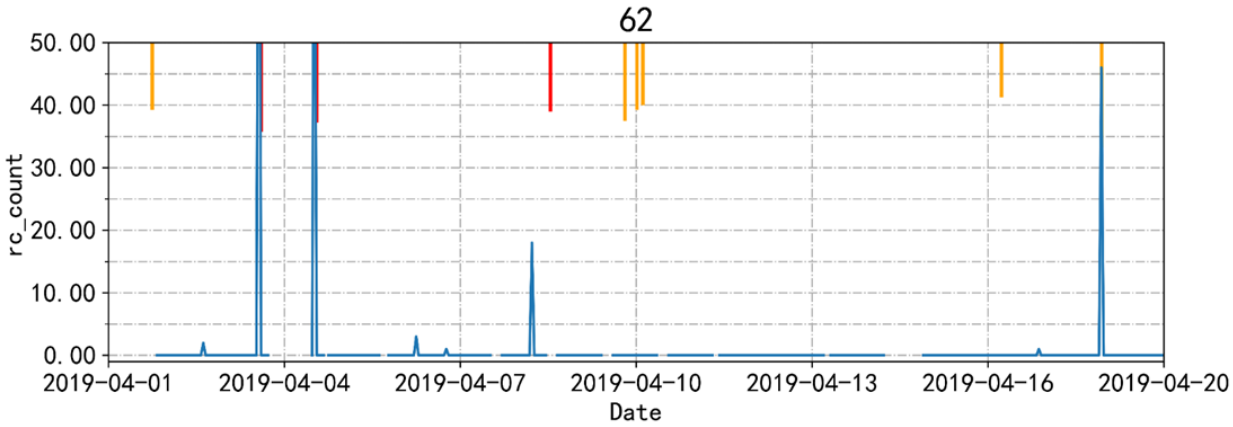

图 3.6 AETA 台站连续采集后数据的映震效果图

3.5 本章小结

本章首先分析了 AETA 台站每个月进行数据传输和存储的成本，进而阐明了抽样采集方案的设计初衷。抽样采集在降低数据传输和存储成本的同时，又会存在丢失数据信息等问题，不满足数据连续性的需求和断裂带检测的需求，因此提出了一种连续采集的方案。同时，为了在实现数据的连续采集的同时，不会产生过大的数据传输和存储的成本，本章又通过分析比较多种无损压缩算法的优点和不足，提出了一种适合 AETA 原始数据压缩处理的无损压缩算法，并最终实现了原始数据的连续采集和压缩处理，满足了数据的连续性，并在 AETA 系统的映震效果上得到了很好的体现。

第四章　AETA 地声原始波形异常信号的分析和特征提取

正如第二章介绍所说，AETA 多分量地震监测系统是一套大区域，高密度布设的用于观测地震前兆异常的综合体系，其主要目标是最终实现对地震三要素的预测。而 AETA 采集到的原始数据，是一种典型的由传感器采集到的时序数据。这种与时间关联的数据具有数据量大、单个数据点信息含量低等特点。AETA 的采集频率为低频采样 500Hz，全频采样 30kHz。如此高的采样率使得每一秒钟都有大量的数据点的上传，如此巨大量的数据既不利于数据波形的可视化，也不方便进行数据分析工作。因此，从原始数据中提取有效的特征数据是一项必要的数据分析工作。

4.1　原始数据三种基础特征的介绍和分析

在系统设计之初，为了减少服务器的计算压力，保证特征数据计算的实时性，对原始数据提取了三种简单的特征数据，分别是均值、振铃计数以及峰值频率。

4.1.1 三种基础特征值的介绍

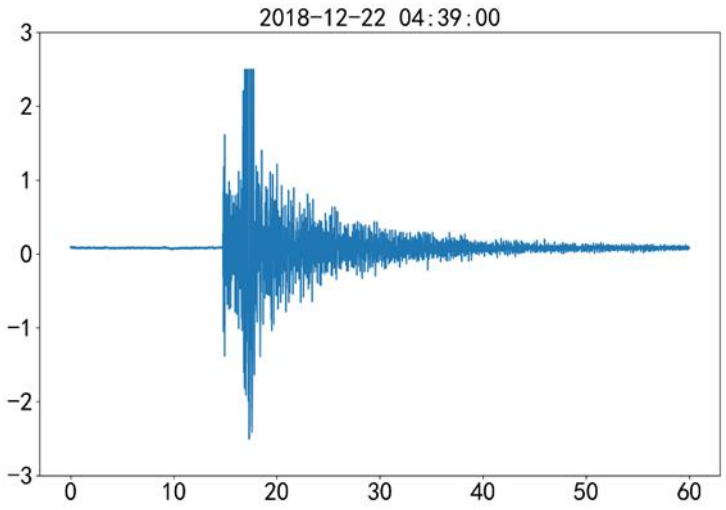

图 4.1 AETA 地声原始数据波形图

（1）均值： 如图 4.1 所示为 AETA 地声原始数据波形，可以看到地声原始数据为有符号数值，且以 0 值为中心上下波动。因此简单的求和取均值的计算方法会因为正负相消的问题而得到一系列绝对值与 0 相近的数据，而丢失了原始数据波形中本身的

幅值信息。为了规避这个问题，均值特征值的计算方法采用的是绝对值求均值的方法：首先设定固定的特征值颗粒度，以颗粒度大小作为时间窗口的长度截取特征数据；然后对每个时间窗口内的原始数据取绝对值，再对绝对值序列求和取平均，得到的均值即为该时间窗口所对应的特征数值。

（2）振铃计数：振铃计数通过计算单位时间窗口内，原始数据波形由上而下和由下而上穿过零点的次数而得。

（3）峰值频率：峰值频率首先是通过快速傅里叶变换将原始数据的频域进行离散化，进而得到离散的峰值序列，然后取其中的峰值作为该时间窗口内原始数据的峰值频率特征值。可以表达出原始数据频域的相关信息。

这三种基础的特征值，有效的解决的原始数据因为数据量太过庞大而无法可视化以及不方便分析利用的问题。在很长的一段时间内，作为了 AETA 多分量地震监测系统数据分析系统的主要研究和实验对象。尤其是基于均值的一些数据分析工作取得了很好的地震前兆异常检测的效果。

4.1.2 基于电磁均值特征的数据分析工作成果

电磁的均值特征一直是数据分析工作的重点研究对象，也取得了许多很好的研究成果：

（1）基于电磁均值的滑动主成分分析方法提取 SRSS 波异常

这种方法是选取了电磁均值呈现 SRSS 波形的台站的电磁均值为实验对象，通过滑动主成分分析的方法提取异常值。该方法以 2017 年 8 月 8 日九寨沟的 7 级地震作为实验对象，取得了很好的实验效果。实验选取的数据的时间范围为 2017 年 7 月 9 日到 2017 年 9 月 7 日，电磁均值处理之前呈现如图 4.2 所示的 SRSS 波形。

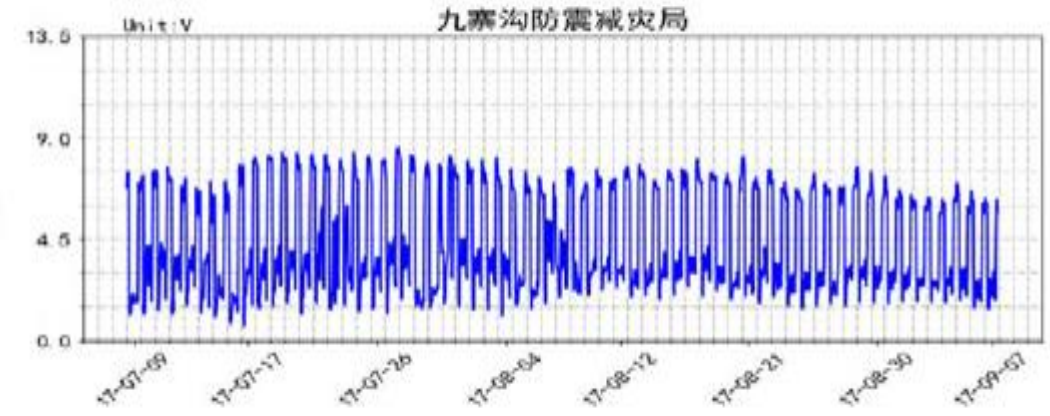

图 4.2 SRSS 波形示意图

通过滑动主成分分析的方法提取这部分数据的异常值，得到实验结果如图 4.3 所示：

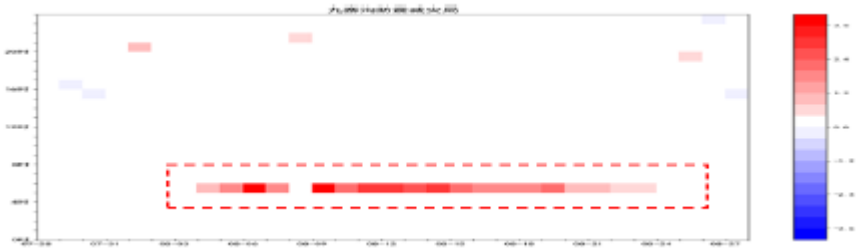

图 4.3 滑动主成分分析法效果图

实验结果显示，八月八日九寨沟发生七级地震的前后几天，九寨沟 AETA 地震监测台站出现了明显的高异常情况。高异常条带开始于震前四天，之后异常值持续明显；而在地震发生之后，异常条带渐渐消失。实验结果验证了基于 SRSS 波的滑动主成分分析方法提取异常值方法的有效性。

（2）分形维数算法提取电磁均值异常值

分形作为非线性特征研究领域的重要分支[77-79]，在 AETA 电磁均值的异常提取中也取得了较好的成果。算法思想是假设 AETA 电磁数据是一个复杂的信号，包含了地震相关数据、人类活动以及其他自然因素的干扰信息。通过分形维数可以描述数据的混沌程度，尤其适用于一维时序数据。通过在时间域计算分形维数的 Higuchi Fractal Dimension(HFD)算法，计算得到 AETA 台站每天电磁均值特征的分形维数，进而生成分形维数序列描述数据的异常值变化情况。

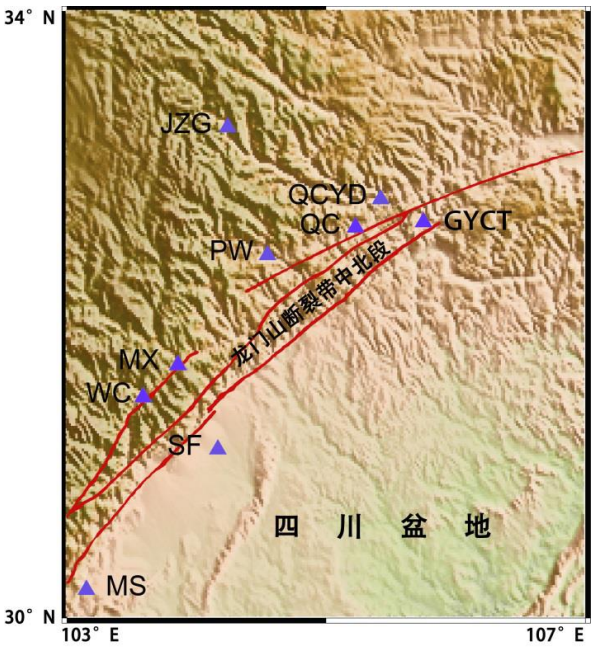

图 4.4 9 个 AETA 台站分布图

以四川省龙门山断裂带北区的 9 个 AETA 观测台站为实验对象，台站分布如图 4.4 所示，具体台站信息如表 4.1 所示：

表 4.1 9 个 AETA 台站位置信息表

台站编号	台站名称	台站简称	经纬度
43	青川县防震减灾局	QC	105.23°E，32.59°N
90	茂县测点	MX	103.85°E，31.69°N
91	汶川防震减灾局	WC	103.59°E，31.48°N
99	什邡市防震减灾局	SF	104.16°E，31.13°N
115	广元市朝天区东溪河台	GYCT	105.75°E，32.63°N
116	平武县防震减灾局	PW	104.55°E，32.41°N
121	九寨沟防震减灾局	JZG	104.24°E，33.25°N
124	名山区安吉村	MS	103.15°E，30.19°N
141	青川县姚渡观测站	QCYD	105.42°E，32.78°N

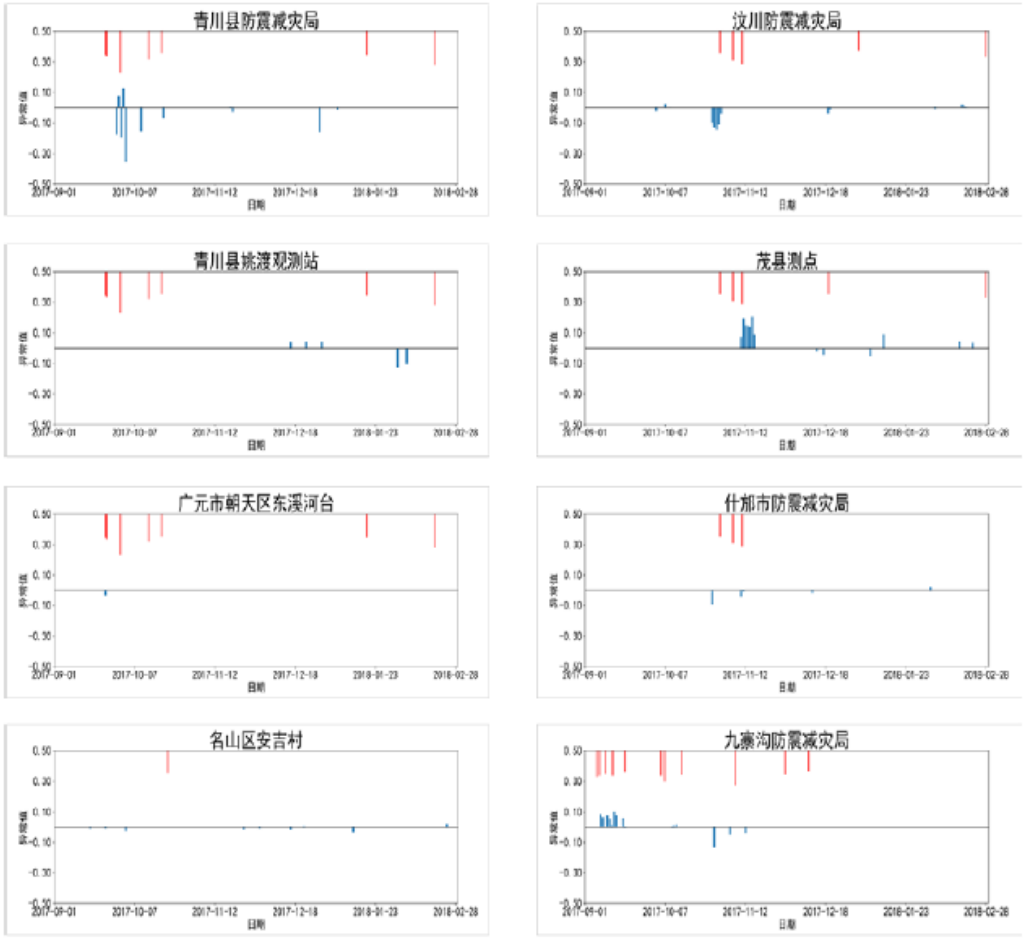

图 4.5 分形维数算法实验效果图

为了将算法提取的异常值与地震事件做可视化对比，图中用红线的线段表示震中距在 55 千米以内并且超过 2 级的地震以及震中距在 110 千米之内并且超过 4 级的地震事件，其中线段长度代表震级的大小。实验效果如图 4.5 所示：

上图显示，在 2017 年 9 月 1 日到 2018 年 2 月 28 日期间，实验选取的 8 个 AETA 台站总共出现的异常次数为 30，其中出现在地震事件前 15 天或者震后 5 天内的异常次数为 19。表明了基于分形维数提取的 AETA 数据异常特征与地震时间具有较好的对应效果。

4.1.3 地声均值特征

相对于基于电磁均值所取得的一系列数据分析研究成果，地声均值的研究遇到了比较大的困难。如图 4.6 所示为九寨沟台站在 2018 年 9 月 12 日 5.3 级地震发生时为期一周的地声均值波形，可以发现地声均值灵敏度较差，几乎成一条平滑的直线，没有显示出与地震相关的异常波形。

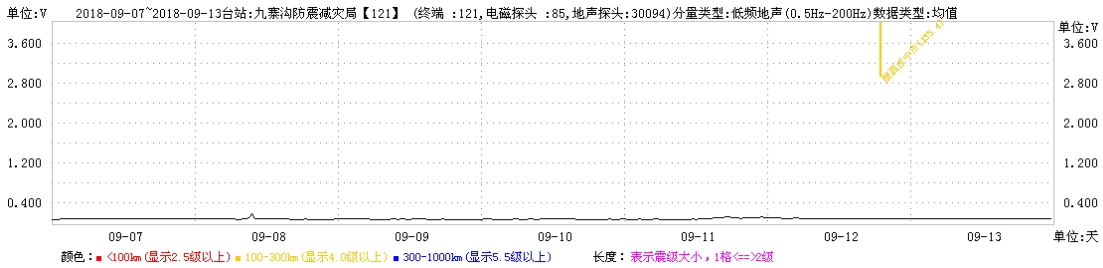

图 4.6 AETA 地声均值波形示意图

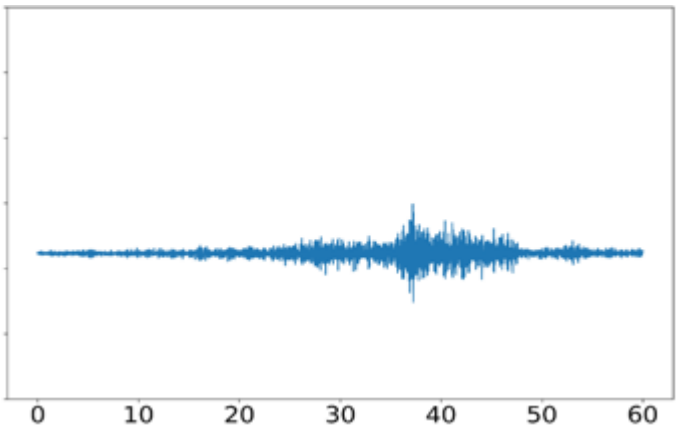

图 4.7 AETA 原始地声异常模式的波形示意图

但是观察地声原始数据波形可以发现，如图 4.7 所示，AETA 原始地声数据会带有不正常的高频模式和高幅值模式。然而从 AETA 原始数据提取特征数据的过程中，简单的均值特征或抽样特征并不能捕获到这些只持续几秒的异常信号。因此本文提出一种

针对性的模式识别算法，从原始数据出发，识别提取出图 4.7 所示的地声异常波形信号以及能够描述这些异常波形信号的特征值。

4.2 基于原始地声数据的异常特征提取

AETA 地声原始数据的规格为：采样频率 500HZ，每三分钟提取一分钟的数据上传并存储，即一分钟的数据量为 30000 条，一小时包含 20 个不连续的分钟的数据。

基于 AETA 原始地声数据的异常检测主要分为如下 3 步：

1）设计模式识别算法，通过算法将地声信号中的如图 1 的异常信号提取出来。

2）以分钟为颗粒度，将模式识别算法处理后的原始地声数据转化为特征数据。

3）将分钟颗粒度的特征数据进一步压缩为小时颗粒度的数据，画出数据波形，与地震事件作对比。

4.2.1 模式识别算法

1）以前一天的数据$\{x_1, x_2, \cdots, x_n\}$为背景，计算背景均值：

$$\bar{x} = \frac{\sum_1^n x_i}{n} \tag{4.1}$$

2）将当天的数据与前一步求得的背景均值做差值得到$\{d_1, d_2, \cdots, d_n\}$，对差值计算均值。

$$\bar{d} = \frac{\sum_1^n d_i}{n} \tag{4.2}$$

3）将$\{d_1, d_2, \cdots, d_n\}$与$\bar{d}$做比较，设立倍数阈值 k，其中 k≥1，如果$d_1 \geq k * \bar{d}$，则第 m 个点即为模式识别算法所要提取到的异常点。

4.2.2 转化为分钟颗粒度数据

通过上一步的模式识别算法，每个数据可以被识别为正常和异常两种模式，然后做如下时序数据处理：

（1）如果两个异常数据之间的间隔小于阈值，则判定量这两个异常数据为同一段异常数据，否则，则分属两段分离的异常数据。

（2）如果一段异常数据点的长度超过阈值，则判定识别到异常地声信号，所属的这一分钟的信号数据标为 1；否则，该分钟的信号数据标为 0，表示这一分钟的时间窗口内不存在异常地声信号。

至此，一个台站每天的原始地声数据从$24 \times (60/3) \times 30000$的规格，转化成$24 \times (60/3)$的由 0 和 1 组成的数据规格，即将每分钟的 30000 个点转化为了一个点，其中 0 表示该分钟的地声数据无异常，1 表示该分钟的数据存在异常地声信号。

4.2.3 转化为小时颗粒度数据

4.2.2 中所得到的由 0 和 1 组成的数据并不能很好的将原始地声波形中的异常信号可视化，本文通过求得每小时的由 0 和 1 组成的数据中 1 所占的比例，将分钟颗粒度的 01 时序数据转化为小时为颗粒度的时序数据，数据值的大小范围为 0~1 的浮点数。然后将得到的以小时为颗粒度的时序特征数据以波形图的形式呈现，并于地震事件做对比，即可直观观测到 AETA 地声数据与地震之间的关系。

4.3 基于 AETA 原始地声数据的异常特征提取算法的应用

为了规避实验的偶然性，验证算法的泛化性，实验分别选取了 AETA 多分量地震监测系统布设于广东汕头、四川珙县、四川九寨沟、河北唐山和北京平谷的五个不同地区的台站的原始地声数据。地震事件的筛选条件如下：1）与台站的直线距离在 100km 内、并且震级大于 3 级的地震，标为红色；2）与台站的直线距离在 100km 到 300km 范围内、并且震级大于 4 级的地震，标为黄色。

实验效果图的作图规则如下：

（1）横坐标为时间坐标轴；纵坐标为原始地声数据提取的异常特征值的大小；

（2）蓝色曲线代表算法所得异常特征值随时间的变化情况；

（3）上方的红色和黄色竖线代表对应时间发生的地震事件，竖线长短代表地震震级大小，每格长度对应两个地震等级。

实验结果如图 4.8 所示，每张子图上方的红色和黄色竖线代表对应时间和震级的地震事件，绿色和红色的地震序号分别表示震前是否有出现对应的高异常特征值。

2018 年珙县地震台周围符合条件的地震事件总共有 6 个。从图 4.8 第一张子图中可以看到，这六个地震事件中有五个地震在震前 1-2 月都有显著的高异常特征值，尤其是发生于 12 月 16 日、距离台站直线距离 22.2km 的 5.7 级地震前一个月的时间窗口中有全年最高的异常特征值波动。

图 4.8 第二张子图为广东汕头地震台的实验结果图，其中符合条件的地震事件为发生发生于台湾海峡的 6.2 级地震及其余震，如表 2 所示。而在这次 6.2 级地震前后一个月的时间窗口，汕头地震台地声信号的异常特征值有显著的高值波动情况。而在全年其他时间，特征值处于较低水平。

图4.8第三张子图为北京平谷地震台的地震事件和地声异常特征值的对应效果图。该台站敏感性较低，但是依旧可以看出在唯一一个符合条件的地震事件发生前后，台站的异常特征值出现了不同于全年其他时期的起伏波动情况。

图 4.8 第四张子图为唐山滦南地震台的实验结果图。符合条件的地震事件为

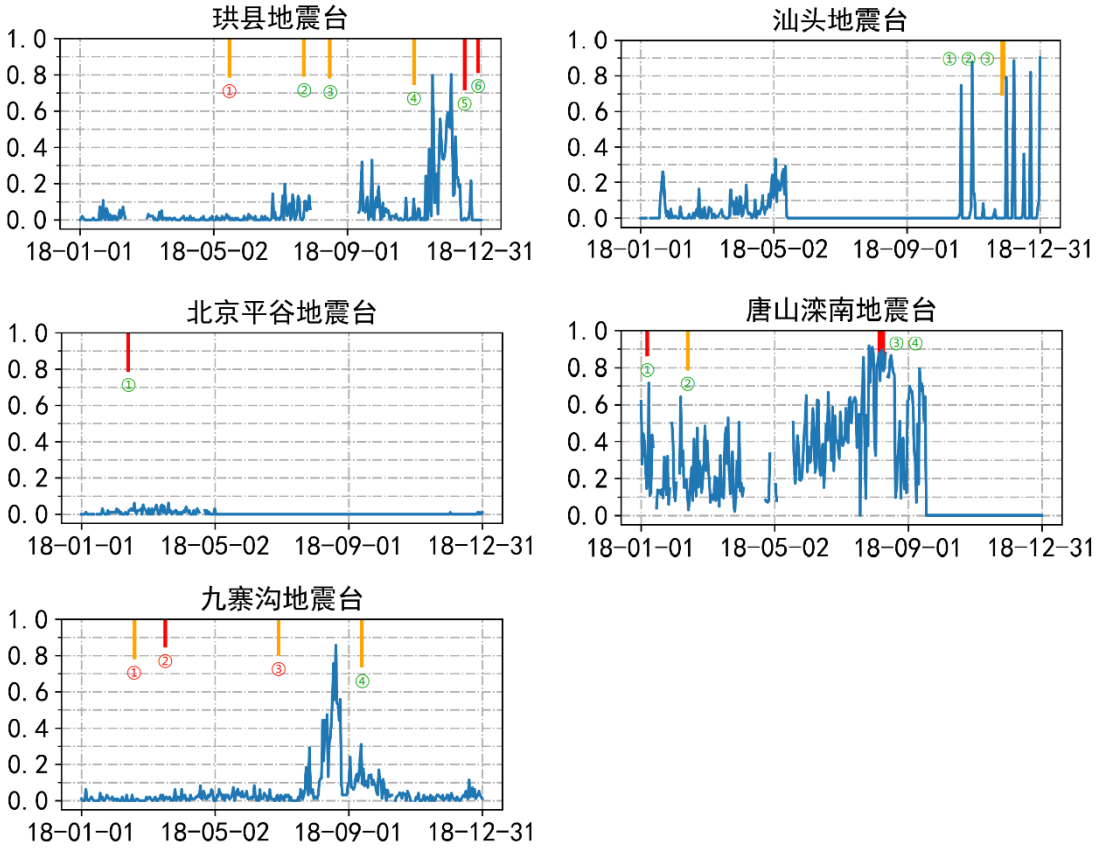

图 4.8 5 个 AETA 台站的地声特征波形图

4 个，其中 3 和 4 号地震为距离台站最近的连续地震事件，恰好对应了全年的地声异常特征值的峰值信号，而 1 号和 2 号地震事件前一个月窗口也有较高的异常特征值波动信号出现。

图 4.8 最后一张子图为四川九寨沟地震台的实验结果图。图中显示，全年地声信号只有九月份前后出现一次显著的高异常特征值，恰好对应了 9 月 12 号发生的震级最高的 5.7 级地震。而其他符合条件的地震事件并没有得到高异常值的对应。

本文将具体台站和地震事件的经纬度和时间信息整理于表 4.2 至表 4.6；台站的相对位置分布见图 4.9，其中蓝色五角星代表五个台站的位置，绿色实心圆代表震前有高异常值的地震事件，红色实心圆代表震前无高异常值的地震事件。

本文选取的五个分布在不同地区的 AETA 台站在 2018 年共计出现 18 次地震事件，实验结果统计显示其中多达 14 次地震在震前出现了显著的高异常特征值，召回率为 77.8%；而未召回的几次地震有着距离台站远、震级不大、集中分布于某一个台站等特点。而在无地震事件的时间窗口范围内，五个台站基本都没有出现明显高于其他时间的高异常特征值的情况。

表 4.2 珙县地震台（104.79°E，28.38°N）周围地震事件

序号	异常	震中位置	经度/°E	纬度/°N	震中距/km	震级	时间
1	×	四川雅安市石棉县	102.28	29.19	260.6	4.3	2018-05-16
2	✓	四川内江市威远县	104.55	29.55	132.2	4.2	2018-07-23
3	✓	贵州毕节市威宁县	104	27.43	131.1	4.4	2018-08-15
4	✓	四川凉山州西昌市	102.08	27.7	276.5	5.1	2018-10-31
5	✓	四川宜宾市兴文县	104.94	28.23	22.2	5.7	2018-12-16
6	✓	四川宜宾市筠连县	104.65	28.15	29.0	3.8	2018-12-28

表 4.3 汕头地震台（116.66°E，23.4°N）周围地震事件

序号	异常	震中位置	经度/°E	纬度/°N	震中距/km	震级	时间
1	✓	台湾海峡	118.6	23.28	198.5	6.2	2018-11-26
2	✓	台湾海峡	118.65	23.21	204.3	4.5	2018-11-26
3	✓	台湾海峡	118.65	23.33	203.2	4.3	2018-11-27

表 4.4 北京平谷地震台（117.03°E，40.14°N）周围地震事件

序号	异常	震中位置	经度/°E	纬度/°N	震中距/km	震级	时间
1	✓	河北廊坊市永清县	116.67	39.37	90.9	4.3	2018-02-12

表 4.5 唐山滦南地震台（118.66°E，39.51°N）周围地震事件

序号	异常	震中位置	经度/°E	纬度/°N	震中距/km	震级	时间
1	✓	河北唐山市滦县	118.46	39.66	23.9	3	2018-01-06
2	✓	河北廊坊市永清县	116.67	39.37	171.6	4.3	2018-02-12
3	✓	河北唐山市古冶区	118.45	39.78	35.0	3.3	2018-08-05
4	✓	河北秦皇岛市卢龙	118.96	39.93	53.3	3	2018-08-09

表 4.6 九寨沟地震台（104.25°E，33.26°N）周围地震事件

序号	异常	震中位置	经度/°E	纬度/°N	震中距/km	震级	时间
1	×	四川广元市青川县	105.02	32.29	129.7	4.4	2018-02-18
2	×	四川阿坝州九寨沟	103.79	33.24	42.8	3.1	2018-03-18
3	×	四川绵阳市平武县	104.57	32.17	124.8	4	2018-06-29
4	✓	陕西汉中市宁强县	105.69	32.75	145.7	5.3	2018-09-12

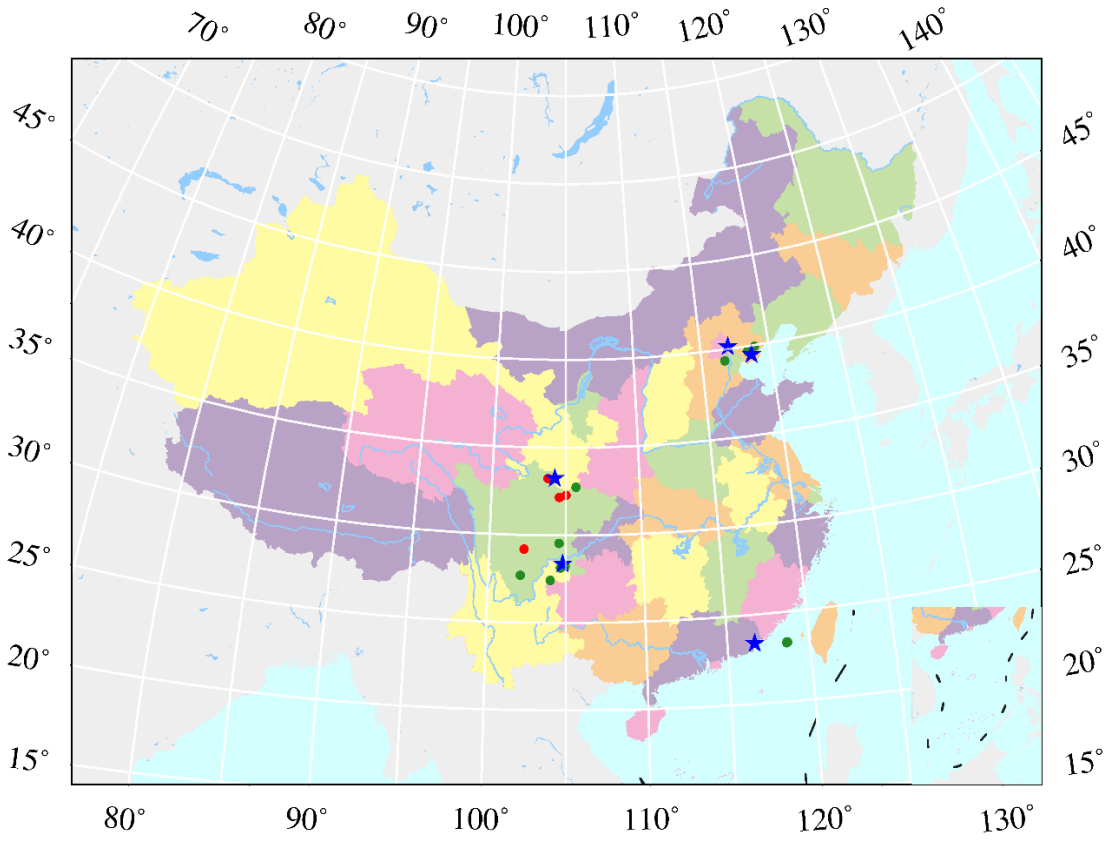

图 4.9　5 个 AETA 台站及周围地震事件的位置分布图

4.4　本章小结

本章实现了对 AETA 原始地声数据的异常模式识别和特征提取。把地声探头采集到的连续的、巨大量的原始地声数据，通过一种模式识别算法，自动化地识别出原始地声波形中的异常模式，并且转化为可以描述地声异常程度大小的特征值。

本章首先介绍了 AETA 系统原有的均值、振铃计数和峰值频率三种基础特征值，分析了基础特征值的优势和不足。总结比较了电磁和地声基于三种基础特征值的地震前兆异常检测的研究现状，发现地声的基础特征值没有发挥出应有的作用。通过将地声原始数据进行合理的可视化，并与地震事件做对比，发现原始地
声波形中存在与地震具有较高相关性的异常模式，根据这些异常模式的特点，设计了一种可以自动化识别出原始地声波形中的异常模式的模式识别算法。通过模式识别算法识别出地声原始数据中的异常模式，标为异常点，最后通过颗粒度转化将原始数据转化为描述地声波形异常值程度大小的特征数据。为了验证模式识别算法和提取的特征值的泛化性和效果，选取了位于全国不同地区的多个台站的地声原始数据作为研究对象。实验结果表明，基于地声原始数据新提取的特征值，与地震事件具有很好的对应效果，并且有明显的前兆性，能够在地震发生之前显现高异常值。

第五章　地震事件的标签转化算法和相关性分析

在经过第三章和第四章对地声原始数据的连续采集、压缩处理、异常模式识别和特征提取之后，AETA 地声原始数据转化成为以小时为颗粒度的描述地声异常程度大小的特征数据。为了评估特征的有效性，本章主要完成了以下两个研究工作：

1.设计了地震事件的标签转化算法，将一个个孤立的地震事件转化为时序数据，以便于将特征与地震进行相关性计算和分析。

2.分析比较了多种不同的相关性计算方法，从不同的角度计算基础特征和新提取的特征与地震事件的相关性，验证了特征的有效性。

5.1　地震事件的异常值标签转化算法

为了满足计算特征与地震事件相关性以及建立模型进行预测地震的需求，需要将孤立的地震事件转化为样本标签，从而方便比较不同特征与地震之间的相关性大小关系以及为建立预测模型提供可以学习的标签。由于最终的目标是要预测地震的三要素：地震发生的时间、震级以及地震发生的地点，因此地震事件的标签转化算法需要综合考虑地震的三要素。本节综合考虑了时间间隔、地震震级、地震中心与台站的距离三方面因素分别对数据的影响程度，建立了一个将孤立地震事件量化为数据标签的算法。转化算法是在以下几个假设和猜想的基础上建立的：

1.地震的发生并非偶尔事件，而是在各种自然因素以及人为因素的影响下，不断孕育并最终达到某个临界点而发生。而地震前兆异常的检测工作主要目的就是从地震数据中挖掘发现这些与地震事件高度相关的异常。模型的目的是通过建模将输入的地震数据转化为地震事件的异常值并输出，从而达到提前预测地震三要素的目标。

2.地震发生之前的前兆异常和地震发生之后的异常通过能量体现在监测的地震数据当中，数据的异常起伏大小与地震的能量成正相关。

3.地震孕育过程中积累的能量呈现不断增大的状态，体现在监测到的地震数据中的异常也越来越明显，地震发生之后由于能量的迸发消散而使得地震数据中的异常也趋于平静并最终消失。

4.地震事件对地震数据异常值的影响与距离呈现反相关关系。与震中距离越小的台站采集到的地震数据，异常越明显。与震中距离较大的台站采集到的数据异常较小。

5.1.1 地震震级的标签量化

自 1953 年震级的概念提出以后，地震学研究领域中，可以通过震级来估计地震的

能量，即著名的古登堡-里克特震级-能量关系式：

$$\lg E_s = 1.5M_s + 4.8 \tag{5.1}$$

式中E_s的单位是 J。需要注意的是，利用古登堡-里克特计算震级与能量的关系式是具有一定局限性的，只是一种粗略的计算方法，其中震级与能量呈现一种 e 指数的关系。根据本文建立算法的假设 2 所描述：地震事件造成的数据异常大小与地震的能量呈现正相关关系，再结合计算简单的原则，在不影响相关性计算的条件下，本文建立如下数据异常大小与震级的关系式：

$$\lg E_s = M_s \tag{5.2}$$

其中E_s表示本文需要从数据中挖掘的异常值大小，M_s表示造成数据异常的地震事件的震级。可以这样建立关系式的原因是，并没有改变能量和震级之间 e 指数的关系，也并不会影响后续的相关性计算的结果以及用于预测地震三要素的模型的结构。

5.1.2 时间窗口的标签量化

正如假设一所描述，现阶段的地震学研究认为，地震的发生并不是偶然事件，而是地壳介质在构造动力的作用下，随着应力累积达到破裂极限时所发生的快速强烈的破裂甚至断层错动所致[80]。而地震发生之后，随着能量的急剧崩发，积累的能量也呈指数型衰减。因此，综上考虑，建立如下地震事件发生前后数据异常值与时间窗口之间关系的标签量化算法：

$$E_s = \begin{cases} \dfrac{E}{1 + e^{-k(t+d)}}, t < 0 \\ e^{-kt}E, t \geq 0 \end{cases} \tag{5.3}$$

其中，t 为台站采集的地声数据距离地震发生时间的时间间隔，t 为负值表示数据在地震发生之前的异常值大小，t 为正值表示数据在地震发生之后的异常值大小。E 表示地震发生之时数据的异常值，为最大值。k 和 d 均为人为设定阈值参数，用来控制波形起伏变化的速度，属于超参数。

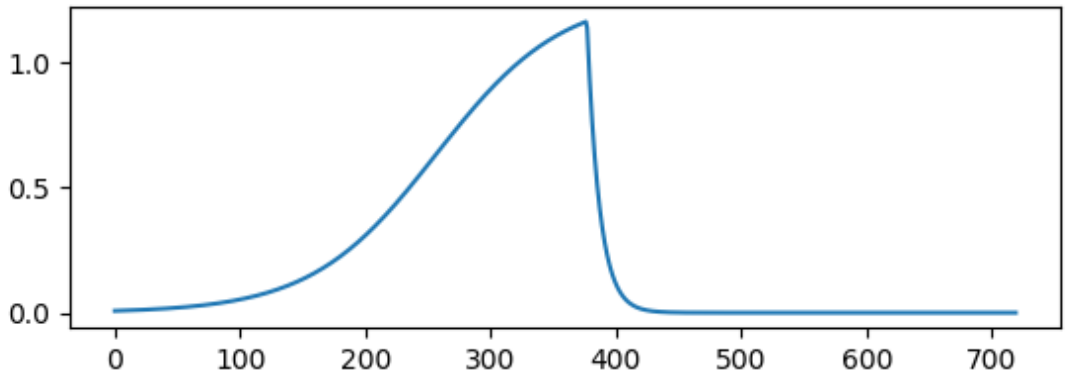

图 5.1　数据异常的大小与距离地震事件的时间间隔的关系示意图

图 5.1 展示了在固定了震级大小和震中距之后，单个地震事件中数据异常程度标签的大小随时间间隔的变化情况。在地震发生之前，由于能量的不断积累逐渐接近临界点，地震数据中异常标签的值与时间间隔也呈现出一种类似于 sigmod 函数的关系，在不断的增大；而在地震发生之后，由于积累的能量瞬间迸发并消散，地震数据中的异常标签的值与时间间隔也呈现出一种指数衰减的关系。

5.1.3 震中距的标签量化

如假设 4 所述，地震的能量和震前异常波动与距离呈现反相关关系，即距离震中越远，受影响的程度就越小。因此建立如下地震数据异常的大小与震中距的关系式：

$$E_s = e^{-ks}E \tag{5.4}$$

其中，s 为采集数据的台站距离地震发生中心的距离，E 为震中的异常值标签大小，k 为阈值参数。

5.1.4 多台站的标签量化

综合 5.1.1-5.1.3 的算法设计，可以建立如下地震数据中的异常大小与地震事件三要素的关系：

假设存在一个孤立的地震事件，震级为M_s，附近有一个 AETA 地震监测台站，与地震中心的距离为 s，那么可以将这个孤立的地震事件转化为一系列与时间 t 相关的标签数值序列：

$$E_s = \begin{cases} \dfrac{e^{-ks}\cdot e^{M_s}}{1+e^{-k(t+d)}}, t < 0 \\ e^{-kt}\cdot e^{-ks}\cdot e^{M_s}, t \geq 0 \end{cases} \tag{5.5}$$

通过上述公式可以将地震事件转化为综合考虑了地震震级，震中距离以及地震发生时间三要素的标签量化时序数据，从而可以计算某个地震事件与该地震事件影响的台站数据提取的特征值的相关性。

但是在实际情况中，每个台站并不是只受一个地震事件的影响，每个数据样本都可能受采集该数据前后发生的多个不同的地震事件的影响。因此在计算特征与地震事件相关性的时候，不能以地震事件为中心，而应该以台站为中心，固定时间窗口的范围，筛选符合条件的该台站周围发生的地震事件，将所有的地震事件的异常值综合考虑，转化为一个描述该台站数据异常的时序数据，从而与该台站数据提取的特征值进行相关性计算以及后续模型的建立。本文采取的方案是，对于每一个台站每一时刻的数据样本，将周围所有符合条件的地震事件转化得到的标签量化数值进行累加。因此在多地震集中发生的区域内，台站的地震事件标签数值整体呈现高幅值状态。具体算法如表 5.1 所示：

表 5.1 地震事件的异常值标签转化算法

算法：地震事件的标签转化算法
确定条件： A：选取的一个 AETA 监测台站 $t_1 \sim t_n$：确定的时间窗口范围 筛选地震： 条件 1：地震发生在时间窗口$t_1 \sim t_n$范围之内 条件 2：地震中心与台站 A 的直线距离在 100km 内、并且震级大于 3 级或者地震中心与台站的直线距离在 100km 到 300km 范围内、并且震级大于 4 级 方法： 步骤 1：按顺序选取符合条件的一个地震 步骤 2：根据震级计算地震发生时震中心的最大能量异常值e^{M_s} 步骤 3：计算该地震与选取台站的距离 s，得到台站在地震发生时的最大能量异常值$e^{-ks} \cdot e^{M_s}$ 步骤 4：根据所选取的时间窗口范围，计算该时间窗口内不同时间点距离地震发生时间的时间间隔 t，计算得到对应时间点的标签数值 $$E_s = \begin{cases} \dfrac{e^{-ks} \cdot e^{M_s}}{1 + e^{-k(t+d)}}, t < 0 \\ e^{-kt} \cdot e^{-ks} \cdot e^{M_s}, t \geq 0 \end{cases}$$ 步骤 5：重复上述操作，将所有地震得到的标签时序数据按照对应时间点累加，最终得到该 AETA 台站在$t_1 \sim t_n$时间窗口内的地震事件经过量化的标签时序数据

由于每个台站某一时刻的数据受周围多个地震事件的影响，因此转化得到的异常值标签序列并不是如图 5.1 的标准单峰波形图，而是多个异常波形图叠加所得，以第四章中选取的 AETA 多分量地震监测系统布设于广东汕头、四川珙县、四川九寨沟、河北唐山和北京平谷的五个不同地区的台站作为实验对象，得到如图 5.2 所示的地震事件标签序列波形图。在一些周围地震频繁发生的时间窗口内，波形整体会呈现高幅值状态，与多地震时期地壳活跃、数据异常明显的情况相吻合。而当台站周围地震稀疏甚至没有地震发生的时候，异常波形幅值也呈现低且平稳的状态。值得注意的是，将图 5.2 与图 4.8 对比可以看出，第四章中由地声原始数据提取得到的地声特征波形图与本章通过转化算法得到的地震事件标签序列波形图具有较高的相关性。

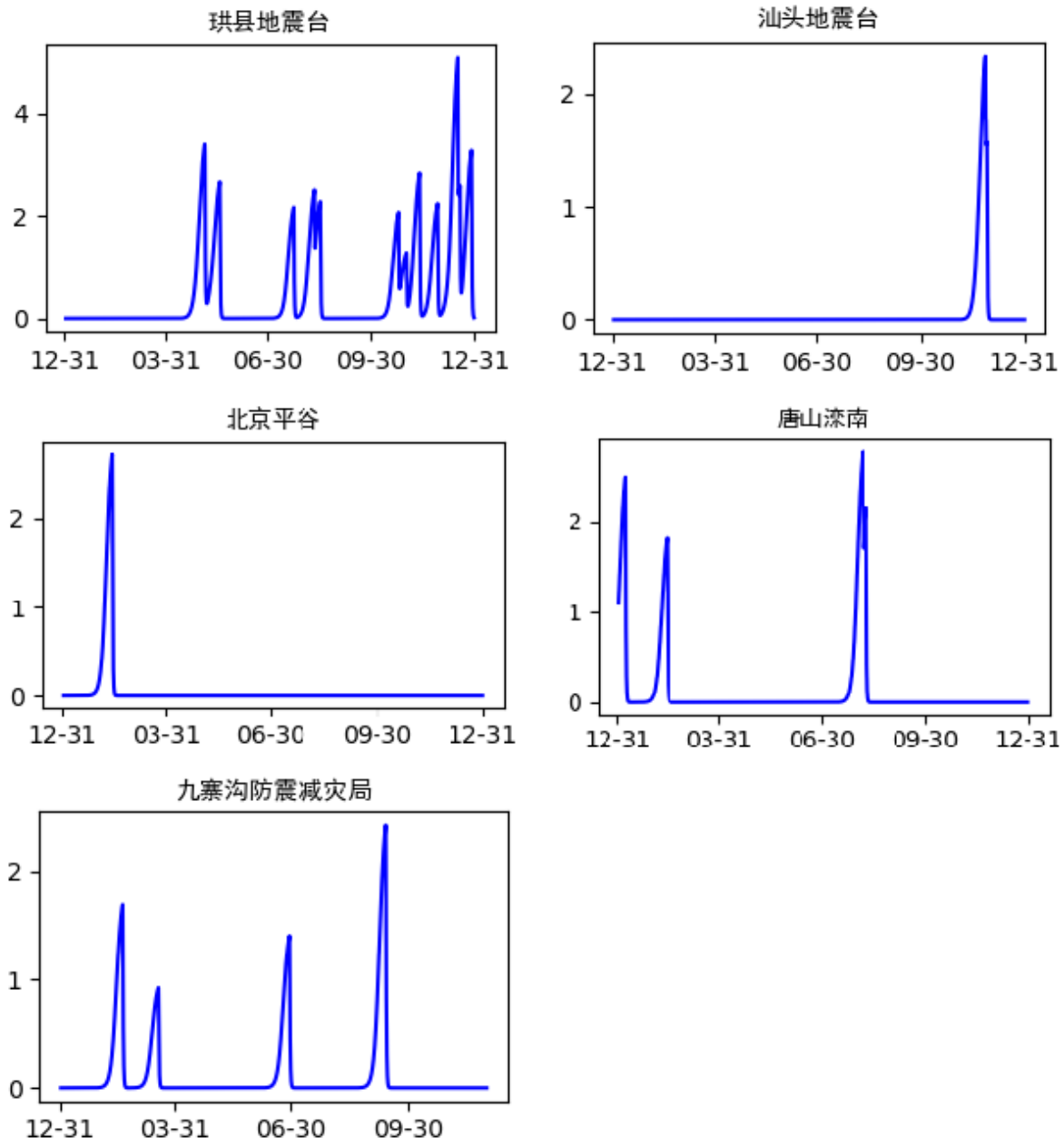

图 5.2　五个 AETA 台站周围地震事件转化的标签序列波形图

5.2　时序数据的相关性分析方法

经过第四章和 5.1 节的模式识别算法、特征提取和地震的标签转化算法本文得到了由 AETA 原始地声数据提取的异常特征值和以台站为中心由周围地震事件转化得到的异常值标签时序数据，为了合理的评估提取的特征值对于地震前兆异常检测的作用，我们需要将特征值与地震事件做相关性计算和分析。

对于两个时序数据的相关性计算方法，主要有皮尔森相关系数、最大信息系数、距离相关系数和基于模型的相关性计算等方法，不同的相关性计算方法有不同的原理、

优势和不足。本文通过调研、分析比较不同的相关性计算方法，从不同的角度计算并分析特征和地震事件的相关性，分析总结第四章提取的新的特征对比原有基础特征的效果差距。

5.2.1 皮尔森相关系数

皮尔森相关系数是一种经典的相关性评判标准。在介绍皮尔森相关系数之前，首先介绍几个统计学中的基本数学概念的计算公式。

（1）数学期望，计算公式为：

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{5.6}$$

（2）方差：

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{5.7}$$

（3）标准差：

$$\mathrm{S} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{5.8}$$

（4）协方差：

$$\mathrm{cov(X,Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{5.9}$$

两个变量的协方差大于零时表示二者的变化趋势一致。而如果两个变量的协方差接近 0 则表示二者之间互不干扰，彼此独立，因为两个独立的随机变量满足：

$$\mathrm{E(X,Y) = E(X)E(Y)} \tag{5.10}$$

协方差用来度量各个维度偏离其均值的程度。然而，协方差的局限性在于只能描述二维空间的情况，对于多维问题需要进行

$$\frac{n!}{(n-2)! * 2} \tag{5.11}$$

次协方差的计算，为了解决这种局限问题，可以选择将多维的数据用矩阵的形式代替。而皮尔森相关系数则是用于解决不同的量纲造成同样的两个量的协方差在数值上表现出很大的差异的问题。

（5）皮尔森相关系数：

$$\mathrm{Corr(x,y)} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} \tag{5.12}$$

皮尔森相关系数的这种计算方式可以有效的消除在计算两个变量相关性的时候，不同量纲对于两个相同的量的影响。

皮尔森相关系数主要有以下几种性质：

（1）有界性：取值范围为-1 到 1

（2）有符号：正负分别表示两个变量之间相关性是正相关还是负相关。

皮尔森相关系数虽然可以衡量两个时序数据之间的相关性，并且可以判断为正相关还是负相关，但是由于他的计算方法和性质，使它具有以下几种比较大的使用局限性：

（1）皮尔森相关系数只能计算出两个变量之间的线性相关性。

（2）两个变量的总体需要满足条件：二者是接近正态的单峰分布。

（3）两个变量的观测值是成对的，每对观测值之间相互独立.

因此，皮尔森相关系数法可以评判特征与地震事件相关性的方法之一，但是需要其他相关性计算方法的补充。

5.2.2 最大信息系数

在介绍最大信息系数之前，首先介绍一个关键概念：熵。熵是用来度量给定概率分布的不确定性。概率分布描述了与特定事件相关的一系列给定结果的概率，其计算公式为：

$$\mathrm{H(X)} = -\sum_{k=1}^{N} P(X=k) log_2 P(X=k) \tag{5.13}$$

概率分布的熵是每个可能结果的概率乘以其对数后的和的负值。对于一个概率事件，如果其结果的分布不确定性越大，那么该事件的熵就越高。因此，当一个事件每个不同的结果发生的概率相同时，则该事件的熵也就最高。即具有均匀分布特性的事件对应的熵最大。

交叉熵是熵的一个拓展概念，它引入了第二个变量的概率分布。

$$\mathrm{H(X,Y)} = -\sum_{k=1}^{N} P(X=k) log_2 P(Y=k) \tag{5.14}$$

两个相同概率分布之间的交叉熵等于其各自单独的熵。但是对于两个不同的概率分布，它们的交叉熵可能跟各自单独的熵有所不同。这种差异，或者叫“散度”可以通过 KL 散度（Kullback-Leibler divergence）量化得出。为了发现变量具有相关性，KL 散度的用途之一是计算两个变量的互信息（MI）。

联合分布和边缘分布乘积之间的散度越大，两个变量之间相关的可能性就越大。两个变量的互信息定义了散度的度量方式。而其中一个重要假设就是概率分布是离散的。那么如何把这些概念应用到连续的概率分布呢？一种方法是量化数据（使变量离散化）。通过分箱算法（bining），将连续的数据点分配对应的离散类别。将两个随机变量转化为散点图的形式，随机分割随机变量转化得到的散点图。统计分割后的每个小散点图

中的落入概率，即可估算出联合概率密度分布。因此，在理想情况下，如果数据量是无限大的，那么可以认为估算出的联合概率密度分布与真实情况的是相等的。所以 MIC 在数据量越大的情况下效果越好。MIC 的计算公式如下：

$$\mathrm{MIC[x;y]} = \max_{|X||Y|<B} \frac{I[X;Y]}{log_2(\min(|X|,|Y|))} \tag{5.15}$$

MIC 主要有以下几种性质：

（1）具有普适性：它不仅可以发现变量间的线性函数关系，还能发现非线性函数关系(指数的，周期的)；不仅能发现函数关系，还能发现非函数关系(比如函数关系的叠加，或者有趣的图形模式)。

（2）具有均衡性：噪声对于函数关系和非函数关系的影响没有很大的差别。

当统计样本足够的情况下，最大信息系数可以变量之间存在的各种相关关系。但是，当零假设不成立时，MIC 的统计就会受到影响，因为一部分研究人员对最大互信息数的统计能力提出了质疑。这个问题会存在于某些数据集中。另一方面，最大信息系数在计算具有时间差的两个时序数据的相关性时具有一定的不适性。

因此，最大互信息数相对皮尔森系数普适性更强，更适用于 AETA 这种非线性、噪声影响较大的数据的相关性计算，可以作为实验中重要的计算方法之一。

5.2.3 距离相关系数

距离相关系数与皮尔逊相关系数有一些相似之处，但是实际上是用一个完全不同的协方差概念来计算的。距离相关系数通过用与距离相类似的量来替代常用的协方差和标准差的概念。距离相关系数不是根据他们与各自平均值的距离来计算两个时序数据序列如何共同变化，而是通过与其他点的距离来估计他们是如何共同变化的，从而能更好地捕捉两个数值序列之间非线性的依赖关系。其主要计算步骤如下：

（1）假设两个时序数据序列 x 和 y 的长度为 N。

（2）对每个序列构建 N×N 的距离矩阵。距离矩阵中每个点表示相应的距离，即行 i 和列 j 的交点表示向量第 i 个元素和第 j 个元素之间的距离：

$$X_{ij} = \|x_i - x_j\|;\ Y_{ij} = \|y_i - y_j\| \tag{5.16}$$

（3）矩阵是双中心的，即对于每个元素，减去了它的行平均值和列平均值，然后再加上整个矩阵的总平均值：

$$\hat{X}_{ij} = X_{ij} - \bar{X}_i - \bar{X}_j + \frac{1}{N^2}\sum_i^N \sum_j^N X_{ij} \tag{5.17}$$

$$\hat{Y}_{ij} = Y_{ij} - \bar{Y}_i - \bar{Y}_j + \frac{1}{N^2}\sum_i^N \sum_j^N Y_{ij} \tag{5.18}$$

（4）在两个双中心矩阵的基础上，将 X 中每个元素的均值乘以 Y 中响应元素的均值，即可计算出距离协方差的平方：

$$\mathrm{Cov}_D^2(x,y)=\frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N}\hat{X}_{ij}\hat{Y}_{ij} \tag{5.19}$$

（5）得到距离协方差之后，可以计算距离的方差（当两个向量相同时，协方差与方差相等）：

$$\mathrm{Var}_D^2(x)=\mathrm{Cov}_D^2(x,x) \tag{5.20}$$

（6）利用上述公式计算距离的相关性：

$$Cor_D(x,y)=\frac{Cov_D(x,y)}{\sigma_D(x)\sigma_d(y)} \tag{5.21}$$

5.3 地声特征与地震事件的相关性计算实验

上一节所述的三种相关系数计算方法，各有所长，互为补充。皮尔逊相关系数在计算两个时序数据的相关性的同时，还可以得到二者之间是正相关还是负相关的关系；距离相关系数与皮尔逊相关系数相类似，但是可以捕捉两个时序数据之间的非线性相关关系，而最大互信息数计算最为复杂，但是效果更加普适可靠，可以发现两个时序数据之间的线性、非线性和函数、非函数的关系。本节以第四章中选择的位于四川、河北、广东等不同地区的五个台站的数据作为实验对象，分别计算这些台站的地声和电磁的均值、振铃计数与地震事件的相关性，并与第四章中基于地声原始数据提取的新特征作比较。

5.3.1 五个台站的地震事件标签转化波形图

首先通过5.1节中提出的转化算法，将AETA多分量地震监测系统布设于广东汕头、四川珙县、四川九寨沟、河北唐山和北京平谷的五个不同地区台站在2018年全年发生的符合条件的地震事件转化为异常值标签序列，作为相关性计算的标签数据。转化波形如图5.2所示。

5.3.2 相关性计算实验结果

将五个台站的原有基础特征值，包括：电磁均值、地声均值、电磁振铃计数和地声振铃计数以及基于地声原始数据提取的新的地声特征值与地震事件转化得到的时序标签数值做相关性计算和分析，采用的相关性计算方法包括了皮尔逊相关系数、距离相关系数和最大互信息数三种算法。实验的计算结果分别列于表 5.1-5.5 和图 5.3-5.7 中：

表 5.1 珙县地震台（104.79°E，28.38°N）特征相关性计算结果

算法 / 特征	皮尔逊相关系数	距离相关系数	最大互信息数 MIC
地声均值	0.055	0.118	0.202
电磁均值	0.055	0.091	0.119
地声振铃计数	0.051	0.071	0.145
电磁振铃计数	0.039	0.051	0.184
地声新特征	0.096	0.142	0.233

表 5.2 汕头地震台（116.66°E，23.4°N）特征相关性计算结果

算法 / 特征	皮尔逊相关系数	距离相关系数	最大互信息数 MIC
地声均值	0.218	0.232	0.433
电磁均值	0.010	0.084	0.166
地声振铃计数	0	0	0
电磁振铃计数	0.222	0.419	0.350
地声新特征	0.048	0.122	0.259

表 5.3 北京平谷地震台（117.03°E，40.14°N）特征相关性计算结果

算法 / 特征	皮尔逊相关系数	距离相关系数	最大互信息数 MIC
地声均值	0.216	0.248	0.328
电磁均值	0.031	0.173	0.248
地声振铃计数	0.036	0.074	0.228
电磁振铃计数	0.014	0.025	0.168
地声新特征	0.244	0.291	0.375

表 5.4 唐山滦南地震台（118.66°E，39.51°N）特征相关性计算结果

算法 / 特征	皮尔逊相关系数	距离相关系数	最大互信息数 MIC
地声均值	0.005	0.125	0.368
电磁均值	0.204	0.215	0.224
地声振铃计数	0.054	0.072	0.137
电磁振铃计数	0.031	0.147	0.389
地声新特征	0.307	0.281	0.442

表 5.5 九寨沟地震台（104.25°E，33.26°N）特征相关性计算结果

算法 / 特征	皮尔逊相关系数	距离相关系数	最大互信息数 MIC
地声均值	0.143	0.179	0.219
电磁均值	0.074	0.098	0.207
地声振铃计数	0.098	0.122	0.239
电磁振铃计数	0.014	0.068	0.207
地声新特征	0.017	0.126	0.223

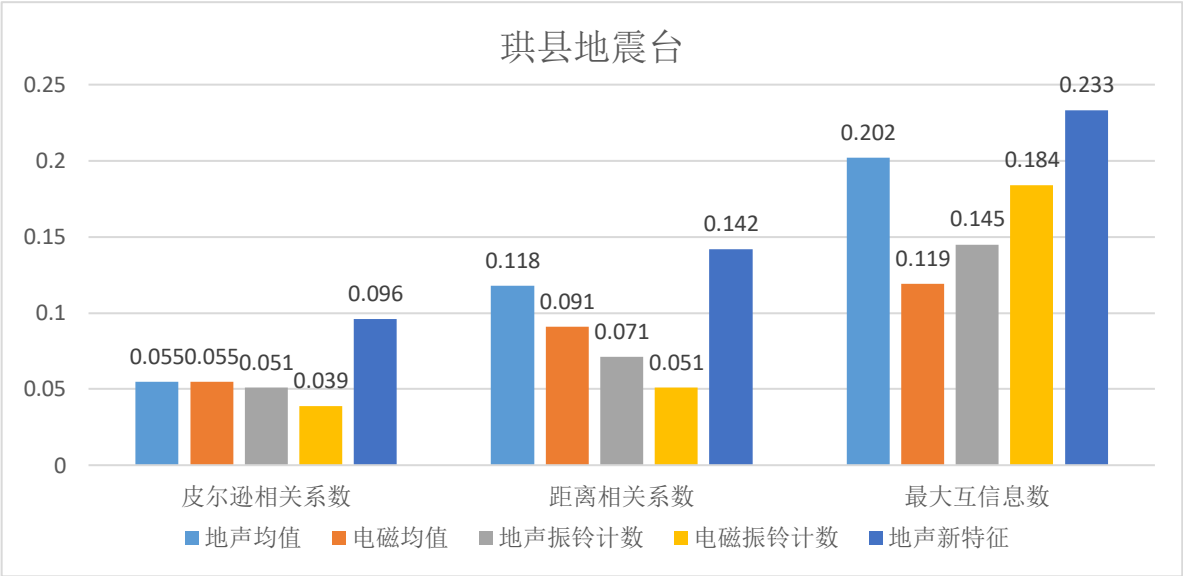

图 5.3　珙县地震台多特征与地震事件的相关性条形图

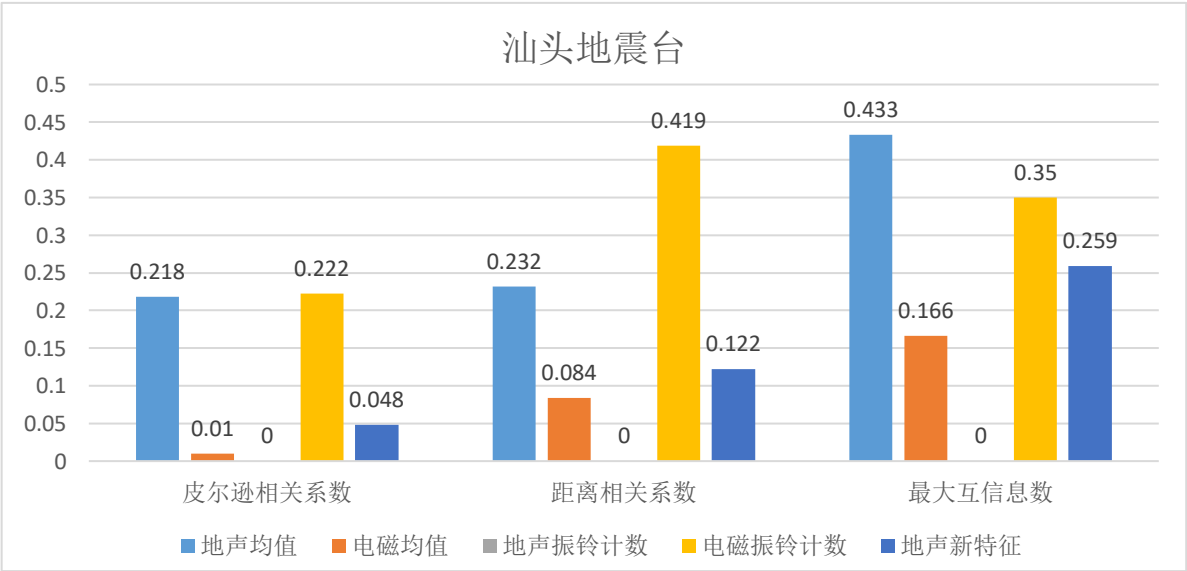

图 5.4　汕头地震台多特征与地震事件的相关性条形图

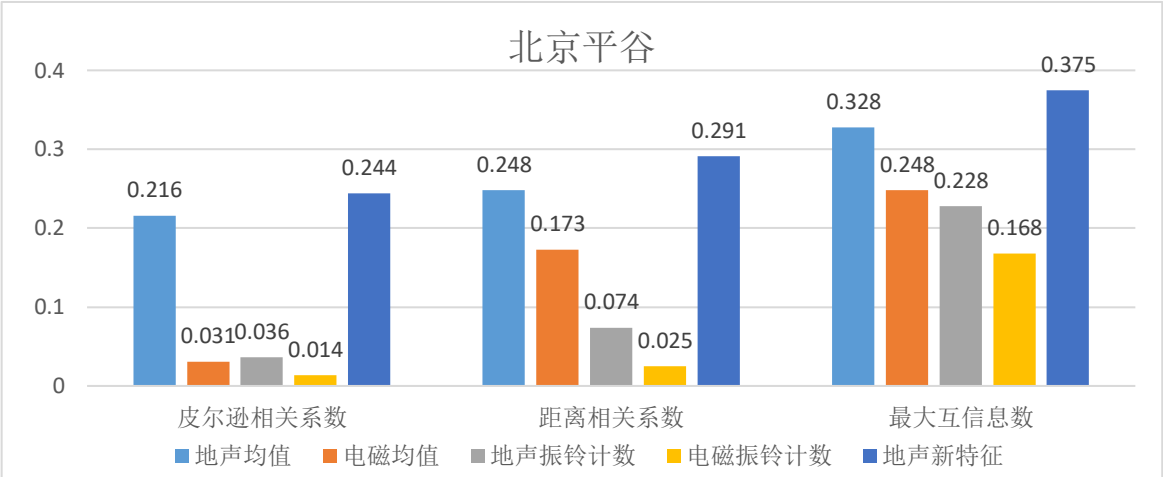

图 5.5 北京平谷地震台多特征与地震事件的相关性条形图

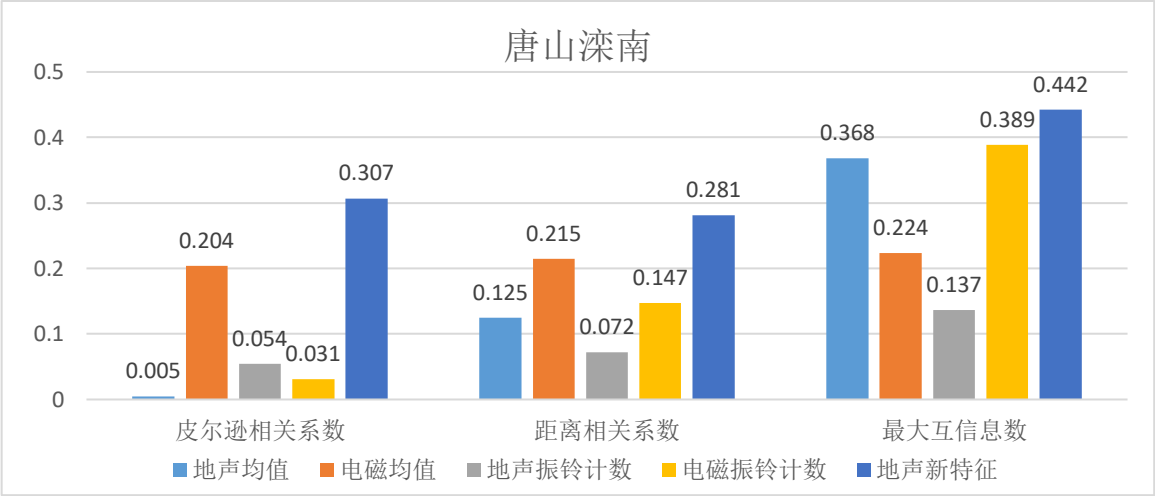

图 5.6 唐山滦南地震台多特征与地震事件的相关性条形图

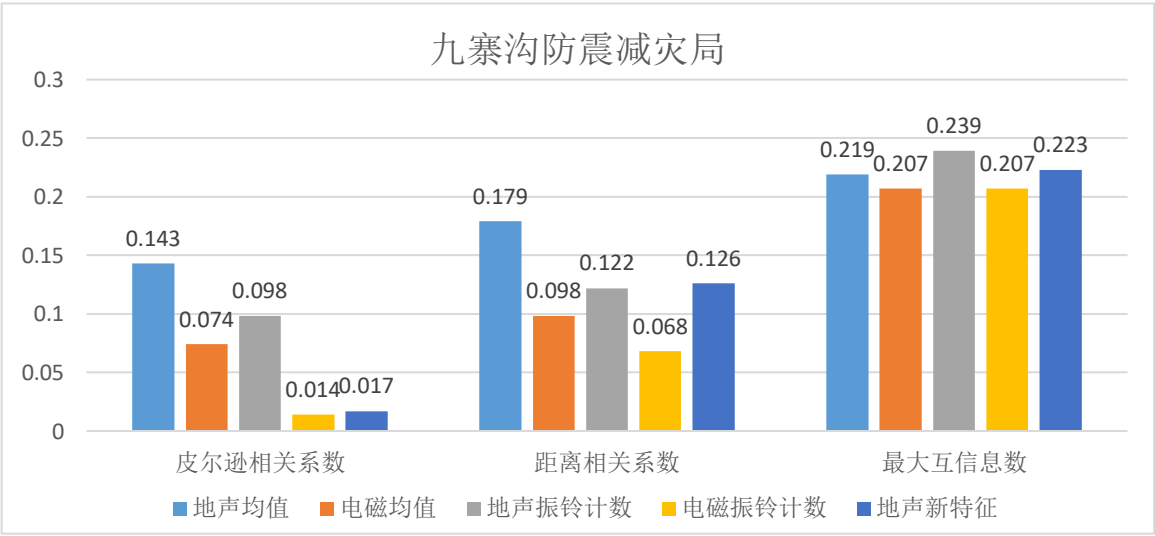

图 5.7 九寨沟防震减灾局多特征与地震事件的相关性条形图

5.3.3 实验结果分析与总结

表 5.1-5.5 和图 5.3-5.7 详细展示了不同特征与地震事件的相关性计算的实验结果。从图中可以看出，第四章中基于原始地声数据提取的新的地声特征，相对于其他特征具有明显的优势。以皮尔逊相关系数为标准，新的地声特征在五个台站中胜出了三个。地声均值特征胜出了两个。以距离相关系数为标准，新的地声特征在五个台站中胜出了三个，地声均值和电磁振铃计数各胜出一次。以最大互信息数为标准，新的地声特征在五个台站中胜出了三个，地声均值和地声振铃计数各胜出一次。综合统计可得，新的地声特征在 15 次相关性结果中胜出了 9 次，地声均值胜出 4 次，电磁振铃计数和地声振铃计数各胜出 1 次。

因此，以地震事件作为效果对比，通过多台站的数据、多算法的相关性计算实验结果可以证明，基于原始地声数据提取的地声特征相对于其他的基础特征而言，对地震前兆异常的检测和地震三要素的预测可以做出更好的贡献。

5.4 本章小结

本章首先设计了一种转化算法，将孤立的地震事件转化成了标签时序数值，并以多个台站周围的地震事件为实验对象，将这些台站的地震事件转化成了标签数值波形，通过与第四章中提取的特征波形做对比，发现二者之间具有较高的相关性。

然后调研了多种不同的相关性计算方法，选取了皮尔逊相关系数、距离相关系数和最大互信息数三种相关性计算方法，以 AETA 多分量地震监测系统布设于广东汕头、四川珙县、四川九寨沟、河北唐山和北京平谷的五个不同地区台站在 2018 年全年的数据作为实验对象，计算了多个基础特征、新提取的地声特征与地震事件的相关性。实验结果表明，相对于原有的地声和电磁的基础特征，第四章中基于原始地声数据提取的地声特征与地震事件具有明显更高的相关性。

第六章 总结和展望

6.1 总结

地震是一种常见的自然灾害。全球每年发生的地震灾害给人民的生命和财产安全带来了严重的损失。为了规避地震灾害带来的巨大损失，国内外许多研究学者致力于地震三要素的短临预测研究。北京大学深圳研究生院集成微系统重点实验室研制的多分量地震监测系统 AETA，在地震短临预测的研究领域已经取得了巨大的进展和成果，采集了丰富的地震数据。本文基于 AETA 系统已有的成果和数据，分别在数据处理终端层面和数据分析层面开展了如下工作：

1. 在数据处理终端层面，介绍了 AETA 系统的组成结构与数据采集流程，通过计算数据传输和存储的成本，从实际需求出发，总结分析了原有数据采集方案的优势与不足，进而提出了对数据采集流程的改进方案：（1）原先每三分钟取一分钟的抽样采集方案导致了数据的不连续性，大大降低了数据的映震效果，丢失了许多有用的信息，并且不能满足 AETA 用于断裂带检测的需求。放弃抽样采集，新的方案将数据的采集方式改为连续采集，并且提出了两种不同的数据压缩处理方案，减少了数据传输和存储的成本。（2）对于绝大多数台站，采用了一种基于平滑连续滤波的有损压缩数据处理方法，既保证了数据的连续性，保留了数据中的对地震前兆异常检测和地震勘探中起重要作用的低频信号，又大大降低了数据传输和存储的成本。（3）作为对照和补充，对于少部分台站，采用了一种新的适用于 AETA 电磁和地声数据的无损压缩算法，保留了全部的原始数据信息，并且降低了数据传输和存储的成本，对地声和电磁原始数据的压缩率分别达到了 0.3 和 0.63 的效果。

2. 在数据分析层面，通过比较地声和电磁的数据分析成果，分析总结了地声均值特征数据的不足，提出了从地声原始数据出发提取与地震前兆异常相关联的新的地声特征以及分析方法：（1）设计了一种可以将孤立的地震事件转化为描述地震异常的标签数值序列的转化算法，通过这种转化算法，实现了地震事件的量化，得到了可以描述地震事件的具体数值标签序列。（2）将 AETA 系统的地声原始数据可视化，挖掘出原始波形中与地震具有较高相关性的异常模式。（3）设计了模式识别算法，可以从地声原始数据中自动化地识别检测出异常波形，并从中提取出新的地声特征。（4）选取了分别布设于广东汕头、四川珙县、四川九寨沟、河北唐山和北京平谷的五个不同地区的 AETA 台站作为实验对象，计算了多个特征与地震事件之间的相关性。实验结果表明，本文基于 AETA 原始地声数据提取的地声特征相对于其他电磁和地声特征，与地震事件之间有

更高的相关性。并且在地震前有明显的特征值升高的情况，能够很好地描述地震前兆异常信息，为地震三要素的预测工作作出了极大的贡献。

6.2 展望

在 AETA 数据分析层面，本文提出了一种将孤立地震事件转化为标签数值的算法，并基于 AETA 原始地声数据提取了与地震事件具有更高相关性的地声特征。地震事件的标签数值化，极大的方便了特征与地震事件的相关性分析等工作，能够更好的以地震三要素的预测目标为导向，研究、提取并分析更多的特征，并且可以完善地震预测模型，因此在未来将进行如下工作：

1. 将以后的基础特征和无监督算法提取的地声和电磁特征与地震事件的标签化数值序列作相关性分析，明确各特征与地震之间的相关性高低，筛选对于地震三要素的预测更有价值的特征。

2. 从 AETA 电磁原始数据出发，将数据可视化，挖掘原始数据波形中与地震具有较高相关性的异常波形模式，建立模式识别算法，从原始数据中挖掘更多的能够描述地震前兆异常信息的特征。

2. 在丰富特征库的同时，抛弃分类模型的思想，以地震事件转化得到的标签数值作为模型预测的目标值，建立回归模型，达到预测地震三要素的目标。

参考文献

[1] 王根龙,张军慧.中国地震灾害防御对策的进展及今后的发展趋势[J].防灾技术高等专科学校学报,2004(02):17-19.

[2] 庄其仁,张渭滨,龚冬梅,王建新.地震前兆地声信号的低频特征[J].华侨大学学报(自然科学版),1999(02):28-31.

[3] 杨瑞华.云南地震地声初探[J].大理师专学报(自然科学版),1997(01):86-87.

[4] 王新安,雍珊珊,徐伯星,梁意文,白志强,安辉耀,张兴,黄继攀,谢峥,林科,何春舅,李秋平.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.

[5] 陈运泰.地震预测——进展、困难与前景[J].地震地磁观测与研究,2007(02):1-24.

[6] Nanjo K Z,Yoshida A. A b map implying the first eastern rupture of the Nankai Trough earthquakes.[J]. Nature communications,2018,9(1).

[7] 张晶,祝意青,武艳强,张希,杨国华.基于大地形变测量的中国大陆中长期强震危险区研究[J].地震,2018,38(01):1-16.

[8] 路鹏,李志雄,陶本藻,李圣强,吴婷,泽仁志玛.震级频度与古登堡-里克特关系式偏离的前兆意义[J].地震,2006(04):1-8.

[9] Jousset Philippe,Reinsch Thomas,Ryberg Trond,Blanck Hanna,Clarke Andy,Aghayev Rufat,Hersir Gylfi P,Henninges Jan,Weber Michael,Krawczyk Charlotte M. Dynamic strain determination using fibre-optic cables allows imaging of seismological and structural features.[J]. Nature communications,2018,9(1).

[10] Marra Giuseppe,Clivati Cecilia,Luckett Richard,Tampellini Anna,Kronjäger Jochen,Wright Louise,Mura Alberto,Levi Filippo,Robinson Stephen,Xuereb André,Baptie Brian,Calonico Davide. Ultrastable laser interferometry for earthquake detection with terrestrial and submarine cables.[J]. Science (New York, N.Y.),2018.

[11] Pisco Marco,Bruno Francesco Antonio,Galluzzo Danilo,Nardone Lucia,Gruca Grzegorz,Rijnveld Niek,Bianco Francesca,Cutolo Antonello,Cusano Andrea. Opto-mechanical lab-on-fibre seismic sensors detected the Norcia earthquake.[J]. Scientific reports,2018,8(1).

[12] 马文娟,刘坚,蔡寅,陈会忠,刘现峰.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(02):835-841.

[13] Skelton A, Andrén M, Kristmannsdóttir H, et al. Changes in groundwater chemistry before two consecutive earthquakes in Iceland[J]. Nature Geoscience, 2014, 7(10): 752–756.

[14] Green H W, Wang-Ping C, Brudzinski M R. Seismic evidence of negligible water carried below 400-km depth in subducting lithosphere[J]. Nature, 2010, 467(7317): 828–831.

[15] Brodsky E E, Thorne L. Geophysics. Recognizing foreshocks from the 1 April 2014 Chile earthquake[J]. Science, 2014, 344(6185): 700–2.

[16] 杨玲英,毛先进.云南地震前兆监测发展[J].国际地震动态,2019(11):42-45+54.

[17] 云南省地震监测志[M]. 地震出版社 ，云南省地震局[编], 2005

[18] 尹祥础,尹灿.非线性系统失稳的前兆与地震预报——响应比理论及其应用[J].中国科学(B 辑 化

学 生命科学 地学),1991(05):512-518.

[19] 陈丽萍,孔祥增,郑之,林新棋,詹晓珊.基于滑动窗口的几何移动平均鞅算法在震前异常分析中的应用[J].计算机应用,2013,33(12):3608-3610.

[20] Alina Marie Hasbi,Mohammed Awad Momani,Mohd Alauddin Mohd Ali,Norbahiah Misran,Kazuo Shiokawa,Yuichi Otsuka,Kiyohumi Yumoto. Ionospheric and geomagnetic disturbances during the 2005 Sumatran earthquakes[J]. Journal of Atmospheric and Solar-Terrestrial Physics,2009,71(17).

[21] M. S. Strigunova,A. M. Shurygin. Sliding-Window Scalar Multiplication of Matrices and Earthquake Prediction[J]. Automation and Remote Control,2004,65(7).

[22] 尹祥础,尹灿.非线性系统失稳的前兆与地震预报——响应比理论及其应用[J].中国科学(B 辑 化学 生命科学 地学),1991(05):512-518.

[23] DeVries Phoebe M R,Viégas Fernanda,Wattenberg Martin,Meade Brendan J. Deep learning of aftershock patterns following large earthquakes.[J]. Nature,2018,560(7720).

[24] 赵纪东,Holtzman B K,Paté A,Paisley J.机器学习发现地震数据中的隐藏信号[J].国际地震动态,2018(07):1.

[25] Holtzman Benjamin K,Paté Arthur,Paisley John,Waldhauser Felix,Repetto Douglas. Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field.[J]. Science advances,2018,4(5).

[26] 李荣安,祝晔,王国治,柴保平.强震宏观前兆对短临预报跟踪的研究[J].东北地震研究,1993(01):69-78.

[27] 曹惠馨,钱书清,吕智.岩石破裂过程中超长波段的电、磁信号和声发射的实验研究[J].地震学报,1994(02):235-241.

[28] 田时秀.地声和地震预报[J].物理,1978(01):58-62+50.

[29] 郑治真. 地声信息工程研究的进展和今后方向[J]. 中国地震, 1989(1): 56–63.

[30] 蒋锦昌, 孙巍, 徐慕玲等. 前兆性地声的衰减特性及生物效应的研究[J]. 地震学报, 1985(2):81–90.

[31] 郑治真. 地震孕育过程中的前兆地声[J]. 地震研究, 1992(2): 193–204.

[32] 陈维升, 李均之, 夏雅琴等. 日本大地震及海啸的早期预测及临震信号[J]. 北京工业大学学报, 2013, 39(8): 1206–1209.

[33] 秦飞, 郑菲, 李均之等. 临震次声异常产生的机理研究[J]. 北京工业大学学报, 2007, 33(1):104–107.

[34] Hill D P, Fischer F G, Lahr K M, et al. Earthquake sounds generated by body-wave ground motion[J].Bull. Seismol. Soc. Am.; (United States), 1976, 66.

[35] Lin T-L, Langston C A. Infrasound from thunder: A natural seismic source[J]. Geophysical Research Letters, 2007, 34(34).

[36] Shani-Kadmiel S, Assink J D, Smets P S M, et al. Seismo-Acoustic Coupled Signals from Earthquakes in Central Italy - Epicentral and Secondary Sources of Infrasound[J]. Geophysical Research Letters, 2018, 45(1).

[37] 张俊兰,周锋.数据压缩的发展历程[J].延安大学学报(自然科学版),2008(03):24-27.

[38] 于翔.数据压缩技术分析[J].青海大学学报(自然科学版),2002(05):52-54.

[39] 桂咏.数据压缩技术及其应用[J].无线电工程,1994(03):64-68.

[40] Xiaoyu Wan. A NEW SIP COMPRESSION MECHANISM BASED ON SIGCOMP IN IMS[C]. 宁波大红鹰学院.Proceedings of 2007 International Symposium on Computer Science and Technology(ISCST'2007).宁波大红鹰学院:2012 中国宁波国际计算机科学与技术学术大会,2007:120-123.

[41] Thomas Boudier,David M. Shotton. Video on the Internet: An Introduction to the Digital Encoding, Compression, and Transmission of Moving Image Data[J]. Journal of Structural Biology,1999,125(2-3).

[42] C. E. Shannon. A Mathematical Theory of Communication[J]. C. E. Shannon,1948,27(4).

[43] David A. Huffman. A method for the construction of minimum-redundancy codes[J]. Resonance,2006,11(2).

[44] Howard Paul G.,Vitter Jeffrey Scott. Analysis of arithmetic coding for data compression[J]. Howard Paul G.;Vitter Jeffrey Scott,1992,28(6).

[45] 邓佩珍.数字图书馆关键技术——数据压缩的原理与方法[J].图书馆学研究,2008(11):35-38.

[46] 吴国清,陈虹,徐小文.基于最优内插预测的科学数据压缩方法[J].计算机科学,2007(08):15-17+44.

[47] M. Parrot, D. Benoist, J. J. Berthelier, et al. The magnetic field experiment IMSC and its data processing onboard DEMETER: Scientific objectives, description and first results[J]. Planetary & Space Science,2006,54(5):441-455.

[48] 汤吉, 赵国泽, 陈小斌等. 地震电磁卫星载荷及现状[J]. 地球物理学进展,2007,22(3):679-686.

[49] Jin Bin Cao, Li Zeng, Feng Zhan, et al. The electromagnetic wave experiment for CSES mission: Search coil magnetometer[J]. Science China Technological Sciences,2018,61(5):1-6.

[50] Chu Baojin, Zhou Xin, Ren Kailiang, et al. A dielectric polymer with high electric energy density and fast discharge speed[J]. Science,2006,313(5785):334-336.

[51] F. U. Liu, HASHIM N. Awanis, Yutie Liu, et al. Progress in the production and modification of PVDF membranes[J]. Journal of Membrane Science,2011,375(1):1-27.

[52] 武艳强, 黄立人. 时间序列处理的新插值方法[J]. 大地测量与地球动力学, 2004(04): 43–47.

[53] 沐守宽, 周伟. 缺失数据处理的期望-极大化算法与马尔可夫蒙特卡洛方法[J]. 心理科学进展,2011, 19(07): 1083–1090.

[54] 刘鹏, 雷蕾, 张雪凤. 缺失数据处理方法的比较研究[J]. 计算机科学, 2004(10): 155-156+174.

[55] 姚休义, 冯志生. 地震磁扰动分析方法研究进展[J]. 地球物理学进展,2018,33(2):511-520.

[56] 李建凯, 汤吉. 主成分分析法和局部互相关追踪法在地震电磁信号提取与分析中的应用[J]. 地震地质,2017,39(3):517-535.

[57] Spiros Papadimitriou, Jimeng Sun, Philip S. Yu. Local Correlation Tracking in Time Series[J]. 2006,

[58] Jinyun Guo, Li Wang, Hongjuan Yu, et al. Impending ionospheric anomaly preceding the Iquique Mw8.2 earthquake in Chile on 2014 April 1[J]. Geophysical Journal International,2018,203(3):1461-1470.

[59] J. Y. Liu, Y. I. Chen, C. H. Chen, et al. Seismoionospheric GPS total electron content anomalies observed before the 12 May 2008 Mw7.9 Wenchuan earthquake[J]. Journal of Geophysical Research

Space Physics,2009,114(A4):231-261.

[60] P. I. Nenovski, M. Pezzopane, L. Ciraolo, et al. Local changes in the total electron content immediately before the 2009 Abruzzo earthquake[J]. Advances in Space Research,2015,55(1):243-258.

[61] K. Hattori. ULF geomagnetic changes associated with large earthquakes[J]. Terrestrial Atmospheric & Oceanic Sciences,2004,15(3):329-360.

[62] T. Higuchi. Approach to an irregular time series on the basis of the fractal theory[J]. Physica D Nonlinear Phenomena,1988,31(2):277-283.

[63] B. Mandelbrot. How long is the coast of britain? Statistical self-similarity and fractional dimension[J]. Science,1967,156(3775):636-638.

[64] Benoit B. Mandelbrot. The fractal geometry of nature.[M].1982.

[65] 张彬彬, 张军华. 地震数据低频信号保护与拓频方法研究[J]. 地球物理学进展: 1-8.

[66] 庄其仁,张渭滨,龚冬梅,王建新.地震前兆地声信号的低频特征[J].华侨大学学报(自然科学版),1999(02):28-31.

[67] 包冬梅.数据压缩算法研究[J].无线互联科技,2019,16(21):112-113.

[68] 吴乐南.数据压缩[M].北京:电子工业出版社,2012.

[69] 高潇.地震波正演波场高效压缩方法[D].合肥:中国科学技术大学,2019.

[70] Ziv J, Lempel A. A Universal Algorithm for Sequential Data Compression[J] IEEE Transactions on Information Theory, 1977, 23(3):337-343.

[71] 王忠效,姜丹.关于 Lempel-Ziv 77 压缩算法及其实现的研究[J].计算机研究与发展,1996(05):329-340.

[72] Ziv J, Lempel A. Compression of Individual Sequences via Variable Rate Coding[J]. IEEE Transactions on Information Theory, 1978, 24(5):530-536.

[73] 王平. LZW 无损压缩算法的实现与研究[J]. 计算机工程, 2002, 28(7) :98-99.

[74] 王平, 茅忠明. LZSS 文本压缩算法实现与研究[j]. 计算机工程, 2001, 27(8):22-24.

[75] Welch. A Technique for High Performance Data Compression[J]. IEEE Computer, 1984, 17(6):8-19.

[76] 许霞, 马光思, 鱼涛. LZW 无损压缩算法的研究与改进[J]. 计算机技术与发展, 2009, 19(4):125-127.

[77] T. Higuchi. Approach to an irregular time series on the basis of the fractal theory[J]. Physica D Nonlinear Phenomena,1988,31(2):277-283.

[78] B. Mandelbrot. How long is the coast of britain? Statistical self-similarity and fractional dimension[J]. Science,1967,156(3775):636-638.

[79] Benoit B. Mandelbrot. The fractal geometry of nature.[M].1982.

[80] 李艳娥,陈学忠.长宁地震前应力变化及震前触发过程探索研究[J].国际地震动态,2019(08):182.

攻读硕士学位期间的科研成果

已录用文章：

[1] **Mengxuan Lv**, Shanshan Yong, Xin'an Wang, A Pattern Recognition Algorithm for AETA Original Geoacoustic Data Anomaly Detection, 2020 IEEE 4th Information Technology, Networking, Electronic & Automation Control Conference (ITNEC 2020), 12-14 Jun.2020.（**EI，已录用**）

已申请专利：

[1] 雍珊珊、王新安、刘聪、李丹、**吕孟轩**、何春舅. 一种用于断裂带勘测的数据处理方法、勘测方法和系统: 中国，201910433583.X，2019-05-23.（已申请）

致谢

三年的研究生涯转瞬即逝，过往的点点滴滴犹在眼前。南燕岁月是我学生旅途的终点，亦是我新一段人生的起点。回顾过往，才发现其中已经浸满了不舍与恍惚。不舍这段岁月的美好、充实，不舍这段路途中青春的同窗笑脸、和蔼的导师面容；恍惚这三年的幸福岁月却是如此飞逝，恍惚自己在不知不觉中已行至终点。犹记得四年前第一次踏进南燕，虽只是一名参加夏令营的青涩少年，但已心属于此。而那时还弥漫着施工烟土的南燕，今已光彩宜人，生气盎然。伴随着南燕的日新月异，当初那个羞涩而紧张的少年，也在科研路上一点点成长，人生路上一步步前行。

这个夏天，南燕花香依旧，而我将成为故人，去往远方。正如离巢的雏燕，已羽翼丰满，毕业的我，也因身边人的谆谆教导和点点陪伴，而不惧风雨，充满希望。

感谢三年来王新安导师对我的教导和帮助。作为一名跨专业的研究生，我在保研面试阶段就遇到了挫折。感谢王老师，第一时间短信联系了我，不仅指出了我的不足和改进方向，还给了我极大的鼓励和信心。进入 IMS 实验室后，王老师给了我们极大的科研方向自主选择的空间，在科研专业知识的传授和指导之余，又从更高的层面对我们的科研生涯，研究生学习和生活方面进行了细致的指导和解惑。作为王老师的学生，我不仅在科研之路上完成了从无到有、从懵懂到成熟的蜕变，更在将来的职业旅途和人生规划上得到了思维方式的成长和进步。正如他一直强调的，AETA 多分量地震监测系统和健康管理系统不仅仅是单纯的科研项目，更是为了服务于人民，致力于解决困扰和危害国家和人民的问题。地震的预测是困难的，更是有价值的，终有一天会被我们攻克的。感谢王老师，成为 IMS 的一员是我研究生三年来最大的幸福。

感谢 AETA 项目组组长雍珊珊博士。正如我们平时称呼她为珊珊姐一样，雍珊珊博士不仅仅是我们的项目组组长，更是我们亲爱的师姐，朋友。除了专业的知识、极强的项目责任心之外，珊珊姐对我们的帮助、关心、爱护更是让我动容。她不仅把我们看作是项目组的一员，更是看作了 AETA 项目组这个大家庭的家人。在对我们进行科研项目上的指导的同时，珊珊姐还会从我们的研究生涯和将来的职业规划上给我们细心的指导，关心我们的生活和心理压力。感谢珊珊姐，她的关心、爱护让我们更加积极地投入到科研项目中，更加高效而快乐地完成科研任务，提高自己的科研和学习能力。我会永远把珊珊姐当作我最亲爱的师姐和朋友。

感谢王培老师。AETA 地震监测系统的稳定运行和设备布置离不开王培老师的辛苦奔波。我会一直怀念与王培老师的每一次聊天，怀念他讲的祖国山南山北的风土人情与各种知识。感谢何春舅老师和陈红英老师，IMS 这个大家庭的井井有序离不开两位老

师的细心和操劳。我们的每一次活动，每一笔报销甚至最后繁琐复杂的毕业流程都离不开两位老师的细心工作和极强的责任心。

感谢我的师兄师姐们，与他们相处的每一天都充满欢声笑语，与他们一起组成的IMS大家庭是我最幸福的守护。玩时他们是我最好的伙伴，不存在一丝一毫的隔阂，是师兄师姐，更是最好最亲密的朋友；学时他们是我的前辈，不仅给予我科研上的帮助，更是我在职业规划和求职路上的先驱者，领路人。感谢他们，给了我最宝贵的建议和指导，给了我最大的信心和希望。

感谢同届的小伙伴、小可爱们。与他们一起奋斗、一起成长的日子是如此的幸福、不舍。我们的相伴充满欢声笑语，我们一起奋斗的岁月硕果累累，我们的未来充满希望，我们的友谊地久天长。虽然分别是篇章的终曲，但是相聚会是永远的期冀。我相信在未来的日子里，我们一定可以常聚首，再言欢。

感谢实验室的师弟师妹们，在他们的身上仿佛看到了我们昨天的身影。我们即将毕业离去，他们也将会在科研路上继续前行，取得更大的成就。

最后，感谢我的家人，谢谢你们在远方给我的关心和支持。

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