# 版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

地震灾害是对人类社会造成损失最严重的灾害之一。长久以来，国内外学者为减轻地震带来的损失，在地震预测的研究上做出了巨大努力。然而目前对地震的发震时间、震中、震级即地震三要素的预测水平依然较低。因此，北京大学深圳地震监测预测技术研究中心研发了 AETA 系统，希望利用 AETA 系统的密集台站、多种分量等特点，对电磁扰动和地声等方面的地震前兆信号进行大密度、长期观测，为地震预测技术做出贡献。

本文从 AETA 系统的观测数据出发，利用统计和机器学习相关理论方法，建立了短临地震风险预测模型，对地震三要素进行预测。本文完成的工作如下：

1．对 AETA 系统数据进行预处理，包括缺失数据处理，时域数据分析以及频域特征分析。

2. 讨论了 AETA 系统数据的特征空间生成方式，分别对单个台站的一元时间序列、多个台站的多元时间序列、研究区域内地震事件频率等方面提取特征，使用滑动窗口生成了基于地理区域划分的特征空间。为了验证生成特征的有效性，利用关联分析相关的理论方法对特征有效性进行评估，并根据评估结果选择了 Mag_SRSS_num、Mag_SVD_max 等 15 个特征进行建模。

3．经算法对比后，选择梯度提升树算法分别对地震时间、震级和震中进行建模预测。为了验证模型效果，本文选取东经 97°E~ 109°E, 北纬 25°N~ 35°N 区域内的 186 个地震事件(>Ms3.0)以及 38 个 AETA 台站在 2017 年 7 月 1 日至 2019 年 3 月 1 日期间的数据作为实验对象进行数据实验。通过交叉验证得到当地震时间窗口 n=7 时，模型 AUC 指标达到最高（0.75）。在 n=7 条件下，模型对无地震、3~4 级地震、4~7 级地震的查准率分别达到 0.50,0.53,0.68；查全率分别达到 0.67, 0.48, 0.63。为了对地震震中进行建模预测，本文按 1°经度和 1°纬度将研究区域进一步细分，每个小区域生成独立样本，从而可以在 111Km*97Km 的区域预测发震可能性。在同样的实验数据下，得到模型对于无地震事件、3-4 级地震、4-7 级地震的查全率分别为 0.84，0.59，0.19；查准率分别为 0.81，0.63，0.20。

本文基于 AETA 系统数据提出了短临地震风险预测模型的建立方法。实验结果表明，本文提出的特征空间生成方法和模型建立方法对于 AETA 数据预测地震三要素问题的解决具有一定意义。

关键词：多分量地震监测系统 AETA，地震数据分析，梯度提升树

# Research on the Risk Prediction Model of Short-impending Earthquakes Based on AETA Data

Bohang Li (Microelectronics and Solid-State Electronics)

Directed by Xing Zhang and Xin'an Wang

## ABSTRACT

Earthquake disaster is one of the most serious disasters to human society. For a long time, scholars at home and abroad have made great efforts in the research of earthquake prediction in order to avoid the losses caused by earthquake disasters. However, at present, the prediction level of the three factors of earthquake occurrence time, epicenter and magnitude is still low. Therefore, Peking University has developed the AETA system. It is hoped that the high-density and long-term observation of seismic precursor signals such as electromagnetic disturbances and geoacoustic signals will be made by using the characteristics of AETA system, such as dense stations and multiple components, so as to make further contributions to seismic prediction technology.

Starting from the data of AETA system and using statistical machine learning theory, this paper establishes a short-term and impending earthquake risk prediction model to predict the three elements of earthquake. Firstly, the data of AETA system are preprocessed, including missing data processing, time domain data analysis and frequency domain feature analysis. Next, the method of generating feature space of AETA system data is discussed. The feature space based on geographic region division is generated by sliding window, which extracts features from single station's time series, multi-station's time series and frequency of earthquake events in the study area. In order to verify the validity of the generated features, the correlation analysis theory is used to evaluate the validity of the features. According to the evaluation results, 15 features such as Mag_SRSS_num and Mag_SVD_max are selected to model.

Finally, after comparing the algorithms, the gradient lifting tree algorithm is selected to model and predict the earthquake time, magnitude and earthquake respectively. In order to verify the effectiveness of the model, 186 earthquake events (>Ms3.0) and 38 AETA stations in the east longitude 97° E 109° E and north latitude 25° N 35° N regions were selected as experimental objects for data experiments from July 1, 2017 to March 1, 2019. Through cross-validation, when the time window n = 7, the AUC index of the model reaches the highest (0.75). Under the condition of n=7, the accuracy of the model is 0.50, 0.53, 0.68, and the recall is 0.67, 0.48 and 0.63, respectively, for earthquakes without earthquakes, earthquakes with magnitude 3-4 and earthquakes with magnitude 4-7. In order to model and predict earthquakes, the study area is further subdivided according to 1 degree longitude and 1 degree latitude. Independent samples are generated in each small area, so that the possibility of earthquake occurrence can be predicted in 111Km*97Km area. Under the same

experimental data, the recall rates of the model for earthquake-free events, earthquakes with magnitude 3-4 and earthquakes with magnitude 4-7 are 0.84, 0.59 and 0.19, and the accuracy rates are 0.81, 0.63 and 0.20, respectively.

In this paper, the method of establishing short-term and impending earthquake risk prediction model is discussed based on the data of AETA system. The experimental results show that the method of feature space generation and model building proposed in this paper have certain significance for solving the problem of three elements of earthquake prediction with AETA data.

KEY WORDS: AETA, Earthquake data analysis，Gradient boosting decision tree

# 目录

# 第一章　绪论

## 1.1　研究背景及意义

大地震是人类社会面临的最重大的自然灾害之一，大地震往往具有强大的瞬时破坏力和严重的持续次生灾害，不但造成重大的人员伤亡，还会造成严重的经济损失，给灾民带来了难以抚平的精神创伤和支离破碎的生活环境。据统计，全球地震灾害造成的死亡人数占所有自然灾害导致的死亡人数总和的 54%，在 20 世纪，平均每年约 1.8 万人死于地震灾害，经济损失逾千亿美元[1]。21 世纪以来，由于人口激增、城镇化建设等原因，地震带来的人员伤亡和经济损失有愈演愈烈的势头。2004 年印尼 Ms9.3 级地震引发的海啸导致 30 万人丧生，超过 51 万人受伤，经济损失超过 130 亿美元。2008 年汶川 Ms8.0 级地震造成 6.9 万人丧生，37 万人受伤，经济损失 8452 亿人民币。2010 年海地 Ms7.0 级地震造成 22 万人死亡，19 万人受伤，经济损失 10 亿美元。2011 年日本福岛 Ms9.0 级地震（东日本大地震）造成 1.5 万人死亡，1.36 万亿人民币经济损失，并导致福岛核电站发生核泄漏事故。大地震对人类社会的破坏力令人触目惊心，但是目前为止，世界上还没有一种可以有效对大地震事件进行预测预报的方法。加之我国地处环太平洋地震带和地中海-喜马拉雅地震带之间，是一个地震多发的国家[2]，随着国家的城镇化建设，人口在城市聚集，一旦没有预兆地在大城市附近发生破坏性地震，后果将不可设想。因此，对于地震事件的三要素的预测研究十分必要。

目前已知的地震前兆信号（或可能的地震前兆信号）主要有地面应变加速[3,4]、重力场变化[5,6]、电磁场变化[7-9]、地磁场变化[10-12]、地电阻率变化[13-15]、地下水化学成分变化[16,17]、大气化学成分变化[18,19]等。根据这些前兆的物理意义，本文可以将其归类为应力信号、电相关信号、化学相关信号。现如今，已有大量基于这些信号的地震预测工作。其中，Hasegawa 等[20]在 2011 年东北奥吉地震后观察到应力场的时间变化。地震前压差应力大小为 5–15MPa，应力模式显示了浅层可能的应变累积。震后应力模式显示了主震破裂后的近沟大滑动。Skordas 等[21]在 2011 年 3 月 11 日日本发生的 Mw9.0 东北地震前约 2 个月，在距震中约 135km 的一个测量站观测到的垂直分量中观察到地磁异常变化约 10 天。此观测结果与日本地震活动自然时间分析的独立近期结果一致，表明地震前兆信号确实存在，但是对于不同的地震，前兆信号并不尽相同。因此，传统的地震预测或者地震前兆信号的研究往往是针对某一个震例或者某一种前兆信号进行探究，难以形成一个较为普适的理论。

自从上个世纪八十年代发现了地震电磁扰动信号以来，电磁扰动信号的研究引起了国内外学者的广泛研究[22-24]，根据法拉第电磁感应定律，电磁扰动信号既可以反映

电场变化也可以反映磁场变化情况，可以作为研究电相关信号的很好的观测量。而应力信号由于与地震的直接相关性更高，一直以来都是地震研究者最为看重的物理量之一，通过对区域应力信号的监测，可以为地震预测提供非常重要的前兆信号[25,26]。此外，传统地震台站数量较少，密度较低，随着互联网、物联网、大数据技术的飞速发展，电磁、地声数据的收集相比于其他观测量而言，技术更加容易实现，成本也更低，非常适合大密度布设台站监测。本文完全有能力对地球进行更加精细的监测，从而利用更大量的数据来尝试解决地震预测这一世界难题。因此，AETA 初步选择了电磁扰动和地声作为观测量并且以此衍生出均值、振铃计数、峰值频率等多种分量，在实验场内大量密布设备，并长期观测（至今已经积累 2 年观测数据），希望能利用多种分量、多个台站、多个地震的数据，建立一个较为普适的模型来尝试解决地震预测问题。

## 1.2 国内外研究现状

传统的地震预测方法主要有加载响应理论[27-29]、地震空区理论[30-32]、应力影区理论[33,34]等。这些方法主要是利用历史地震的长周期发生频率与近期发生频率来对未来区域内地震进行预测，根据这种理论只能预测中长期大地震的发生可能性[35,36]；不过中长期的地震预测不但对人类社会作用不大，而且误差往往很大，例如上世纪八十年代，美国科学家预测 1988 年，帕克菲尔德小镇附近会发生 6 级左右地震，但是这个地震直至 2004 年才发生，足足晚了 16 年[37,38]。

对于短临地震预测的研究，目前学界达成共识的是地震前兆确实存在[39,40]，并且于 1996 年确定了 5 项有意义的地震前兆以及 37 项可能的地震前兆[41]。利用传统方法进行的短临地震预测研究主要是从 IASPEI 下属的地震预测委员会专家确定的 5 项有意义的地震前兆和 37 项可能的地震前兆中选取部分前兆特征，将其数值化后与大地震事件作对比[42]。陈学忠等人对水氡含量和电磁波的空间分布非均匀特征进行了异常提取与研究，发现大地震来临前数月水氡含量和电磁波的空间分布非均匀特征均有上升[43]。郝建国等人对准静电场在震前的异常进行了深入详细的研究，他们认为，理论上 30Hz 以下的超低频段电场在地壳中穿透能力强，衰减小，受干扰少，实验也证实了这一理论，低频电场异常多次在大地震前发生[42,44]。传统的短临地震预测方法提供了很多异常特征提取的思路，也很好的总结了过去几十年地震科学家在地震预测领域付出的辛劳和做出的贡献，然而传统方法普遍缺乏预测效果的评价，而且研究对象局限于某一种或者两种物理量[45-47]。这使得研究人员很难对这些前兆信号的准确性有一个客观的认知，而且地震的成因与发生机理非常复杂，乃至至今学界也没有一个统一的定论，不会存在某一个变量单一决定大地震是否发生，因此需要增加研究特征的维度，从多个角度来描述地震前兆，而非仅仅某一两种信号。

近些年，随着机器学习在结构化数据的预测任务中的表现出众，越来越多的地震研究者将目光投向了机器学习，希望利用机器学习的方法来尝试解决地震预测的问题[48-50]。机器学习主要分为监督学习与非监督学习两种。

Morales-Esteban 等人[51]主要利用聚类技术获得时间序列从而对大于等于 4.5 级地震进行预测。首先，地震被分为不同的组别里并且要确定组的数量。然后，当中大型地震发生时，找到相应的时间序列模式。由西班牙地震研究所提供的地震事件数据的结果和利用非参数统计检验进行讨论，这个结果表明此方法获得的结果有良好的表现和较大的显著性。

朱海宁[52]等人使用小波变换和改进的支持向量机这一回归的方法预测中国 2000 年到 2010 年的最大震级。为了进一步准确预测，该文章还引入了小波变换对 M-T 数据（最大震级-年份数据）进行主周期分析。首先，提取 1900 年到 2010 年的最大震级-年份数据。这个数据是一个时间序列，对这个数据进行连续小波变换（采用 morlet 小波），由此得到小波方差图，进而得到该时间序列的主周期为 25 年。据此，在后文 SVM 特征值选取时，以 25 年为时间窗口进行特征提取。然后，以 25 年的地震频次 N，最大震级 M，平均震级 Mbar，折合能量 N* 为一组数据，以 1 年为步长向后平滑得到输入数据。将数据输入到 SVM 算法中做回归分析得到未来一段时间内的最大震级。文章将 1900-2000 年的数据作为已知样本用于训练，2000-2010 年的数据作为未知样本用于检测。最终得到预测的报准率为 93.75%。

2018 年哈佛大学地球与行星科学学院与谷歌联合在 Nature 上发表了一篇利用深度学习进行余震预测的文章[53]，文章使用了 131000 条主震-余震对数据对深度神经网络进行训练，在测试的 30000 条主震-余震对数据中得到了 0.849 的曲线下面积，高于传统的库伦应力变化模型的预测结果（曲线下面积 0.583），这也是深度学习在地震预测工作中应用的里程碑式的文章。

聚类算法、关联规则等被用来研究地震事件序列，进而进行地震的预测。在无监督学习中，基于聚类技术[54-56]和关联规则[57,58]的方法被用于智利、葡萄牙和西班牙等地震高发地区的地震预测上；Mirrashid[59]等人使用基于模糊 C-均值算法的神经模糊推理系统来预测伊朗的地震；晏昱[60]等人利用关联规则方法对中国地区的地震与 DEMETER 卫星数据之间的关系进行挖掘，得到了显著相关的结论。

## 1.3　本文开展的工作

本文基于多分量地震监测系统 AETA 的数据，利用统计学习方法、机器学习算法建立模型，对地震的时间、地点、震级三要素进行预测。为了达成地震三要素预测的目标，本文主要从以下 4 个方面开展研究：

1．总结地震预测问题的研究现状。对现有的地震监测预测方式方法进行介绍，并总结其优劣性，重点介绍了基于统计学习方法和机器学习算法的地震预测研究。

2．AETA 系统电磁扰动原始数据、地声原始数据的分析以及预处理。AETA 系统数据积累量达 18TB，数据中蕴含着丰富的地震信息。本文对原始数据在时域和频域上进行了分析，从时域上，利用 STA-LTA 算法对数据的幅值、均值、方差等特征进行分析；频域上，利用快速傅里叶变换对数据的频谱特性进行分析。在数据预处理方面，为适应算法对数据处理时的要求，提出了 AETA 系统缺失数据的补全方法。

3．研究基于 AETA 系统数据的特征空间生成方法。使用时间序列数据挖掘相关方法，对 AETA 系统单个台站的时间序列数据进行模式识别、序列描述；使用奇异值分解对 AETA 系统某个区域内多个台站的时间序列数据进行分析。基于统计方法，对地震的时间范围、地点范围、震级范围进行研究，并得出样本标注方法。

4．研究建立基于决策树模型的地震三要素预测模型，并设计相应数据进行对比实验，通过 AUC、准确率、查全率等统计指标对模型的有效性进行检验。

## 1.4 论文组织结构

本文一共有六章，各个章节的内容如下：

第一章：论述本文的研究背景和研究意义以及地震预测问题的国内外研究现状。首先，本文分析了地震给人类社会带来了巨大危害，以及预测地震事件的时间、地点、震级三要素的重要性。其次，分析了当前国内外地震预测技术的研究现状，其中主要研究了短临地震预测技术的研究进展以及基于机器学习和统计学习方法的地震预测技术进展。

第二章：首先，介绍多分量地震监测系统 AETA 的数据来源，包括电磁探头、地声探头监测的物理量意义，电磁探头、地声探头数据的处理方式等。其次，介绍本文对 AETA 系统原始数据的分析和处理方法。最后，根据 AETA 系统数据进行震例分析。

第三章：提出了一种基于 AETA 系统数据的特征空间生成方法。首先，基于单个台站数据，对电磁探头、地声探头的一元时间序列进行特征提取，并验证其有效性。然后，基于区域内的多个台站数据，对电磁探头、地声探头的多元时间序列进行特征提取，并验证其有效性。最后，确定样本的标注方法，包括震级范围选取、时间窗口选取以及震中范围选取。

第四章：基于关联分析方法的特征降维。首先，结合本文目标介绍关联分析方法常用的两种算法 Apriori 和 FP-growth 并论述关联分析在本项目中的应用方法。其次，根据 AETA 系统数据生成地震事件频繁项集和非地震事件频繁项集，以此提升模型的准确度和可解释性。

第五章：提出了短临地震风险模型。首先，结合本文目标，分析了常用机器学习方法决策树算法以及支持向量机算法，并研究这两种算法在建模过程中的应用方法。其次，建立模型分别对地震的发震时间、地点、震级进行预测，评估发震风险。

第六章：总结本文所做工作并对未来工作进行展望。

全文的组织结构如图 1.1 所示。

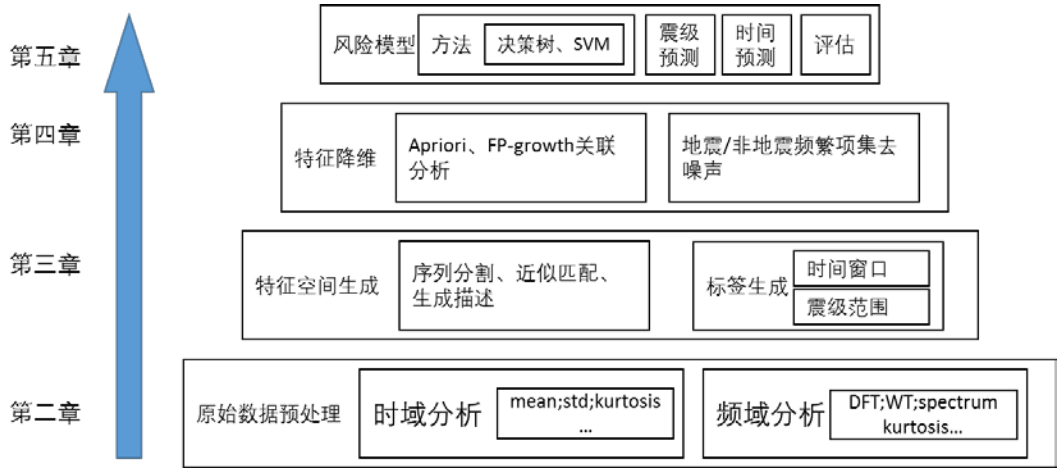



图 1.1　全文组织结构

# 第二章　多分量地震监测系统 AETA 的数据分析

本章主要对 AETA 原始数据介绍、分析，从时域、频域两个方面提取特征序列。AETA 系统有电磁和地声两种探头，其中电磁探头记录电磁扰动信号，按 10KHz 的采样率得到电磁扰动原始数据。地声探头记录地下应力信号，按 50k Hz 的采样率得到地声原始数据。原始数据采样频率比较高，因此数据非常丰富，本文通过滑动均值、滑动方差、傅里叶变换等方法在原始数据基础上提出特征数据并进行滤波。进而，用特征数据与地震事件相对应，得到与地震事件相关性较高的特征。

## 2.1　AETA 系统简介

2012 年起，北京大学深圳研究生院地震监测预测技术研究中心开始研发多分量地震监测系统 AETA[61] (Acoustic and Electromagnetic Testing All in one system)，由数据处理终端[62]、电磁探头[64,66]、地声探头[63,65]、云服务器中间件和数据分析系统组成（如图 2.1 和图 2.2 所示），可以同时监测电磁扰动和地声信号[64,65]。地声传感探头和电磁传感探头一般布设在山洞内或者埋于浅表 2m 处，数据处理终端一般放于机房。数据则通过有线网络或无线网络发送到云服务器中间件进行存储和处理分析。相较于传统地震监测设备，AETA 系统从设备安装层面简化了安装的流程，降低了安装的成本。在设备运营维护方面上也采用远程自动运维的方式提高了设备对环境的适应性。与此同时也尽可能考虑到了对环境干扰因素的消除。


图 2.1　AETA 系统实物图

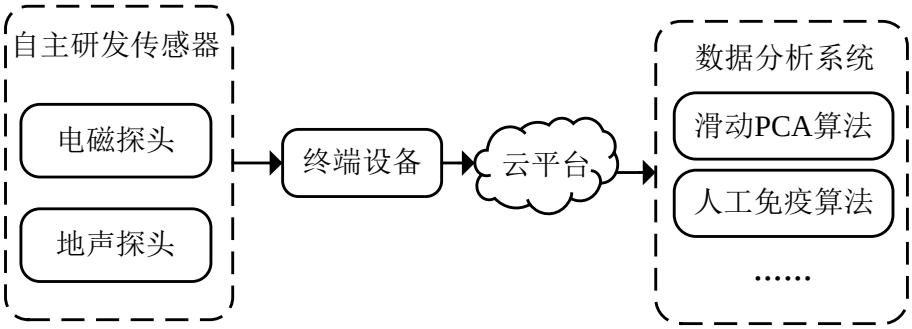



图 2.2　AETA 系统框图

AETA 系统有两个子系统，数据采集系统与数据分析系统：

（1）数据采集系统：包括埋于地下 2 米的电磁扰动探头和地声探头，以及地面机柜中的数据预处理终端；

（2）数据分析系统：包括布设在阿里云上的云服务器、数据分析客户端和数据分析网页。

AETA 系统采集的数据存储在云服务器数据库中，数据可通过云服务器中间件进行 Web 页面可视化，如图 2.3 所示。AETA 数据分析系统的目前的特征数据主要包括以下 3 个部分：

(1) 均值：窗口时间(3min)内原始信号幅值的平均值，可以表征这 3min 内的信号能量，单位为伏特(V)。

(2) 振铃计数：窗口时间(3min)内正向穿越门槛阈值(0)的次数，表征这 3min 内信号的变化频率，检测信号频率异常。

(3) 峰值频率：对窗口时间(3min)内信号进行快速傅里叶变换，其在频域上幅值最大值对应的频率即为该信号的主频率成分，可以表征这 3min 内的信号主频率。


图 2.3　数据查看页面

AETA 系统于 2015 年 8 月完成了第一版设备的小批量试制，快速迭代后于 2016 年 6 月完成了第二版设备 20 套的小批量生产。在中国地震局监测预报司的支持以及四

川、云南、北京地震局协助配合下，两版的设备均进行了现场布设实验。现场试验显示，AETA 系统对当地震例具有较好的捕捉效果，系统灵敏性、稳定性和一致性得到初步验证，2016 年底 AETA 项目与专业硬件服务商深圳卓翼科技达成了深度的合作，AETA 多分量地震监测系统正式定型批量生产，新生产设备进一步布设至中国西南部、首都圈和台湾海峡进行映震试验研究。

截至目前，AETA 系统在中国地震局的支持下，在全国范围内安装约 200 余台，遍布河北、四川、云南、西藏、广东和台湾地区，其中在四川布设设备数量达 93 台，基本覆盖四川全境重点区域。图 2.4 表示的是多分量地震监测系统 AETA 的布设情况。

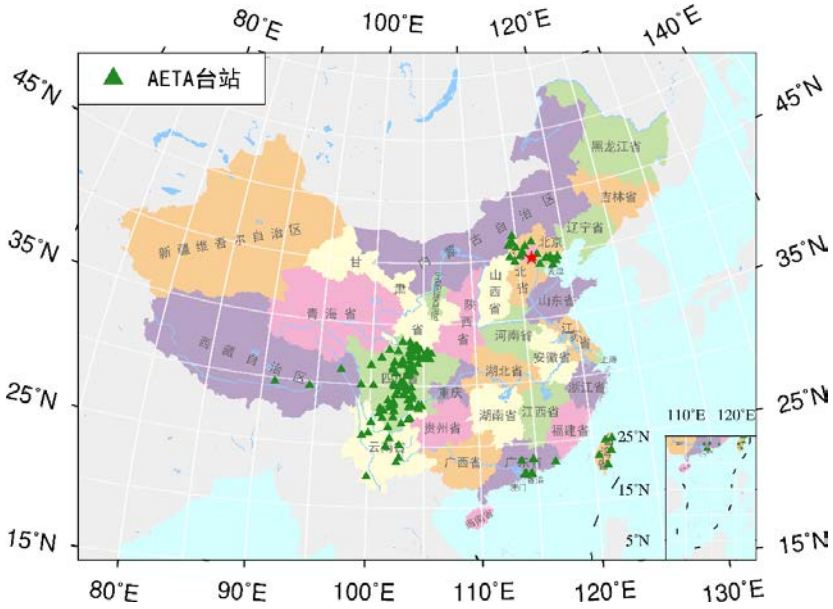



图 2.4　全国 AETA 布设情况

## 2.2　监测物理量介绍

AETA 系统监测的物理量有电磁扰动和地声信号两种，其中电磁扰动信号以 10KHz 频率采样，地声信号以 50KHz 频率采样。Uyeda, Hattori, Han 等人[67-69]的工作中通过实验以及波的衍射透射原理得出，通常与地震相关的电磁信号或机械振动信号来源于震源区以及震源附近区域，频率在零到几百赫兹之间；通常低频波段的电磁波或者振动波传播距离也更远，透射和衍射能力较强。因此，系统也通过滤波得到了低频的电磁扰动信号和地声信号。本小节对系统中的电磁扰动和地声信号与地震的相关性做了全面的分析，为预测模型提供有效的特征。

### 2.2.1　电磁扰动

电磁扰动信号异常在很多地震前夕都出现过，Molchanov、Karalelian 以及汤吉等人的文章[72]中都提到过在地震孕育过程中伴随着不同程度的电磁辐射异常。此外，郝

锦琦[70]和郭自强[71]以及 Warwick、Takeuchi 等还通过岩石破裂实验证实了在岩石破裂过程中有较强的电磁扰动信号产生。至于电磁扰动在地震孕育过程中的产生机理目前还没有定论，学界提出了动电效应(Ivanov, 1939)和微破裂机制(Ogawa, 1985)这两种猜想来解释震前电磁扰动信号的产生机理。动电效应提出岩石破裂后岩石内部的带电荷流体运动改变电磁信号而产生电磁扰动；微破裂机制主要关注岩石破裂后产生的新表面使得电荷分布改变从而改变电磁信号产生电磁扰动。

地震领域研究电磁扰动/电磁信号已经有数十年的历史，前人为本文的研究积累了很多宝贵经验。目前电磁扰动信号与地震之间相关性的研究主要存在的问题和难点在于地震孕育过程中产生的电磁相关信号强度不高，往往还要在地壳中传播数公里到数十公里，穿透不同类型的地理结构到达信号接收台站，所以电磁扰动信号的信噪比往往较低。为了更好的处理电磁信号，一般采用主成分分析法(PCA)、分形分析法以及极化法来对电磁信号进行降噪处理，提高信噪比。

图 2.5　电磁探头实物图

AETA 电磁信号实时采集系统如图 2.5 所示，主要包括信号调理部分、数据采集部分以及数据处理终端交互部分。其中信号调理部分将采集到的 0.1Hz~10kHz 的原始电磁信号进行放大、滤波、A/D 转换后传给数据采集部分；数据采集部分基于微控制器与网络芯片的交互，并将采集到的数据进行打包并发送到远程数据中心；数据处理终端交互部分通过软件实现云服务器远程监控和升级探头。

电磁传感探头：监测频率 0.1Hz~10kHz，强度 0.1~1000nT 较宽动态范围下的低频、超低频电磁波段，灵敏度>20mV/nT@0.1Hz~10KHz，18 位分辨率，采样率为低频 500Hz，全频 30kHz[61]。具体参数如下表所示：

表 2.1　电磁扰动参数表

| 监测对象 | 频段 | 特征数据 | 量纲 |
|---|---|---|---|
| 电磁扰动 | 低频<br>500Hz 以内 | 均值 | 伏：V |
| | | 振铃计数 | 次/每秒：time/s |
| | | 峰值频率 | 赫兹：Hz |
| | 高频<br>30kHz 以内 | 均值 | 伏：V |
| | | 振铃计数 | 次/每秒：time/s |
| | | 峰值频率 | 赫兹：Hz |

### 2.2.2　地下声音

地声信号体现了地下应力变化以及地下岩石振动情况。地下应力和地下岩石振动不论是在地震孕育过程中，还是地震发生时刻都会有不同程度的异常情况出现。对于地声，通常认为有三类产生机制。第一类是岩石的宏观破裂现象，这种现象一般是由地震纵波引起的。第二类是在岩石破裂过程中，岩石中的气体在高温高压等原因下电离，从而在岩石表面放电并产生地声。最后一类是声发射，由于岩石在挤压过程中会发生微破裂，从而产生地声。

AETA 地声信号实时采集系统，如图 2.6 所示，包括地声检测传感单元结构以及信号调理电路结构。

图 2.6　地声探头实物图

地声探头通过将捕获的地声信号传输到压电薄膜传感器，通过压电薄膜将地声信号转化为电信号。然后通过信号调理电路进行处理，再通过电缆传输到数据处理终端，最终上传至云服务器处理后得到地声数据。为了使监测频率包含震前地声信号所涵盖的次声波、 可闻声波、 超声波，采用 PVDF 型薄膜传感器。该传感器的灵敏度为-180dB·V×104/Pa，频率响应范围为 10-3Hz-1MHz，工作温度范围为-50～100 ℃，可以满足 AETA 地震试验场的试验要求。

地声探头的电路由传感器组和信号调理部分组成。其中，传感器组由 3 对、6 条带状压电薄膜传感器构成；信号调理部分由 3 路信号处理链路构成，每条链路依次由带宽为 0Hz~100kHz 的滤波放大电路和 Delta-Sigma 模数转换器（ADC）组成。

该电路结构有以下特点：

（1）检测的范围为低频地声信号(<60Hz)、中频地声信号(60~300Hz)高频地声信号（>300 Hz）以及现有的探头没有检测的超高频地声信号(>1 kHz)，这使得地声信息能够更加完整地保留下来。

（2）该电路的拓扑结构简单，所采用的电路模块分辨率高、功耗低、价格低，极其适用于地声信号的远程监测和地声台网的大面积范围的布设。

（3）该电路结构在传感器、信号调理电路、信号传输部分这 3 个方面都对地声信号进行了备份，很大程度上提高了整个电路系统的可靠性，使得地声探头能适用于可持续地声信号的监测任务。

地声传感探头参数：监测 0.1Hz~50kHz，从次声波、可闻声波再到部分低频超声波波段，灵敏度为 3LSB/pa@0.1Hz~50KHz，18 位分辨率，采样率位低频 500Hz，全频 150kHz[61]。具体参数如下表所示：

表 2.2　地声参数表

| 监测对象 | 频段 | 特征数据 | 量纲 |
|---|---|---|---|
| 地声 | 低频<br>500Hz 以内 | 均值 | 伏：V |
| | | 振铃计数 | 次/每秒：time/s |
| | | 峰值频率 | 赫兹：Hz |
| | 高频<br>150kHz 以内 | 均值 | 伏：V |
| | | 振铃计数 | 次/每秒：time/s |
| | | 峰值频率 | 赫兹：Hz |

## 2.3 原始数据预处理

AETA 系统的原始数据具有采样频率高，数据量大的特点，目前对于原始数据的处理主要从缺失数据处理、时域特征提取和频域特征提取方面入手，试图从原始数据中得到更多地震前兆相关的信息。

### 2.3.1 缺失数据处理

长期监测的时序数据会因为各种原因而存在缺失值，这会给信号的时域频域特性带来很大的影响，因此需要一个策略来处理缺失值。在本文的工作中，以 1 天数据的 10%作为阈值，如果连续缺失 10%以内的数据，本文采用线性插值法来对缺失数据进行补全，如果连续缺失 10%以上的数据，插值补全会引入较大误差，本文会对当天数据做缺失标记，另行处理。

设时间序列为 $TS \in \{(t_i, v_i) \mid i = 0,1,2...N\}$，$t_i$ 为时间序列时间戳，$v_i$ 为时间序列中的第 $i$ 个数据，$i$ 为序列按 $t_i$ 升序排列的序号，N 为序列长度，fs 为时间序列采样频率。本文以一天（24h）的数据为例，fs 为 1/180Hz，也就是以 180s 为采样间隔采样，那么在无缺失的情况下，本文有 $\{(t_i, v_i) \mid i = 0,1,2...480\}$，现有时序 $\{(t_i, v_i) \mid i = 0,1,2...N\}$，遍历 $(t_i, v_i)$，若 $2 \le (t_{i+1} - t_i)/180 \le 48$，说明存在一段连续缺失 10%以下的数据，那么就对 $(t_i, v_i)$ 进行线性插值补全，具体算法 2-1 所示。

表 2.3 线性插值补全算法

**算法 2-1** 线性插值补全算法

**Input** Time series: TS, Parameters: $l$, $fs$, $\delta$ (typical value is 0.1).

**Output** A complete time series, CTS.

1: initialize $\tau \leftarrow \emptyset$, $\gamma \leftarrow \emptyset$

2: **for** each $(t_i, v_i) \in$ TS **do**

3: **if** $2 \le (t_{i+1} - t_i) * fs \le l * \delta$ **then**

4: $\tau \leftarrow \tau$ an array from $t_i$ to $t_{i+1}$ with step $1/fs$

5: $\gamma \leftarrow \gamma$ an array from $v_i$ to $v_{i+1}$ with step $(v_{i+1} - v_i) * fs$

6: **end if**

7: **end for**

8: **return** CTS$\leftarrow(\tau, \gamma)$

### 2.3.2 时域特征提取

提取时域特征，包括均值、幅值 、峭度、方差等。原始数据时间窗口为 1s，如图 2.7 所示。可以看到，电磁扰动信号的原始数据具有频率高、波形随时间变化明显、数

据量大等性质。根据统计，一天一个台站的电磁原始数据量为 100MB，这样的数据量不利于对长期观测数据进行挖掘，因此本文需要在原始数据的基础上对数据进行特征计算，减少数据量，并在此基础上尽可能提升数据质量。

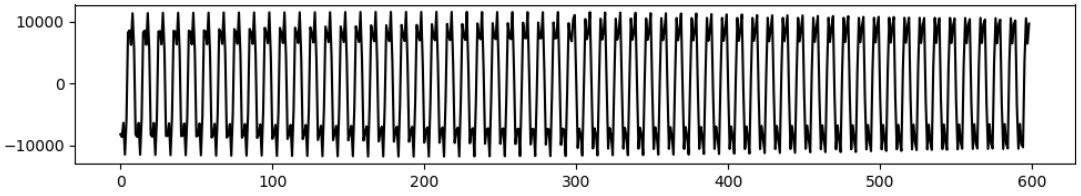



图 2.7　原始数据

首先，对电磁扰动原始数据进行了均值特征处理，也就是滑动平均值处理，对电磁扰动原始数据每 3 分钟计算一次均值，这样一个台站每天产生 480 个数据，极大地减少了数据量，有助于观测数据的长期规律。图 2.8 展示了不同台站不同时段内的均值特征数据。

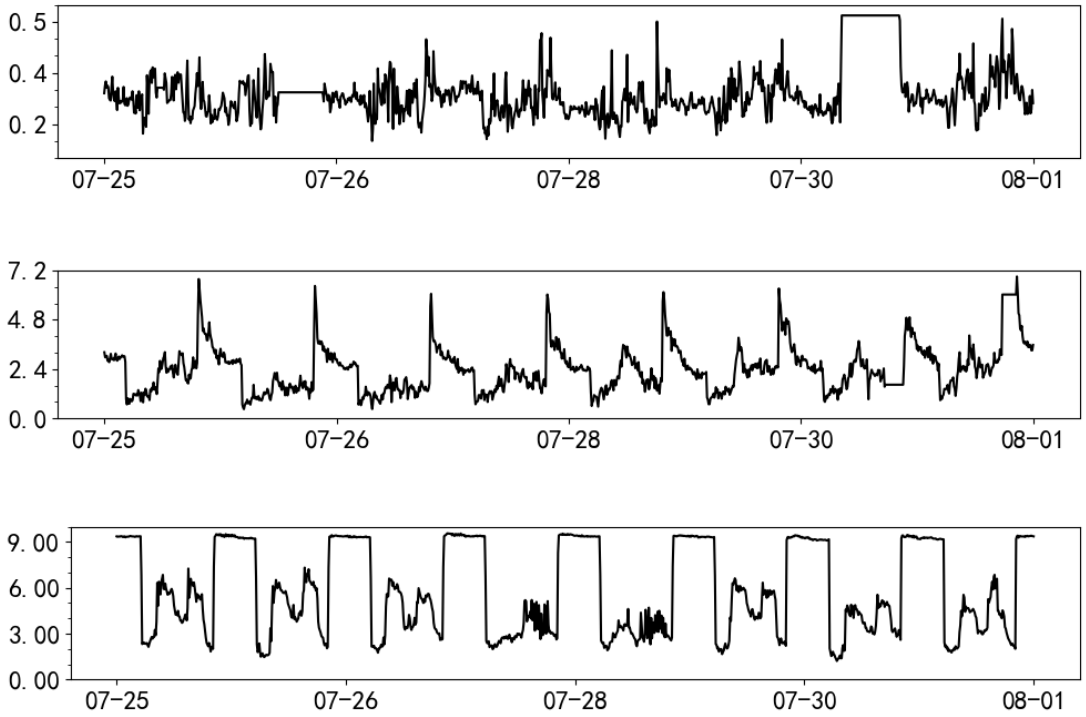
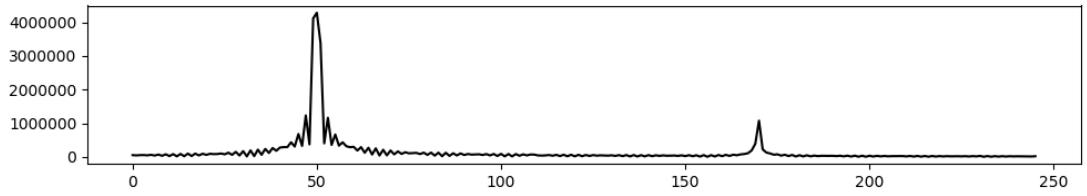



图 2.8　均值特征数据

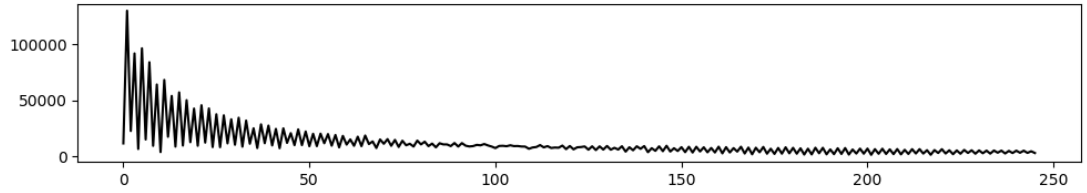
其次，本文对电磁原始数据进行了振铃计数计算，振铃计数是指一定时间内数据穿越 0 值的次数，振铃计数一定程度上反映了原始数据一定时间内的频率大小。

### 2.3.3 频域特征提取

提取频域特征，包括频率、频谱峭度等。以快速傅里叶变换为例，图 2.8 中可以看到这段原始信号中主要频率为 50Hz，次要频率为 170Hz，此外，低频信号能量高于高频信号。这样的频域分析可以提供更丰富的频域特征。




(a) 正常频谱




(b) 低频频谱

图 2.9 原始数据频谱

而有一些台站的原始数据的频谱则更加集中在低频，这种变化通常在时域和特征数据中无法体现出来。图 2.9 展示了九寨沟地震前西昌气象局台站(XC)出现的低频电磁信号，主频率为 7.32Hz。

## 2.4 AETA 观测数据震例分析

在对地震风险进行建模预测之前，对 AETA 系统数据进行相应的震例分析有助于在特征提取和模型建立时对数据有更深刻的认知。因此，本文分别对电磁、地声数据在 AETA 系统开始布设以来的部分典型震例中的可能的前兆信号进行了数据探索和分析。

### 2.4.1 电磁数据震例分析

在电磁数据分析工作中，在长期的数据观测中发现了一些可能与大地震孕育过程相关的现象。其中最为明显的是 SRSS 波，SRSS 波是 AETA 项目组根据长期系统数据以及地震事件的观测发现的一种电磁现象，主要表现为电磁数据幅值随日升日落现象同步变化。

在 2017 年 8 月 8 日九寨沟 7 级地震前后，AETA 系统中，九寨沟防震减灾局（JZG）、松潘防震减灾局（SP）、沐川防震减灾局（MC）、冕宁防震减灾局（MN）、峨眉山防震减灾局（EMS）以及唐山滦南防震减灾局（TSLN）这 6 个台站低频电磁均值信号均出现了 SRSS 波形，如图 2.10 所示。

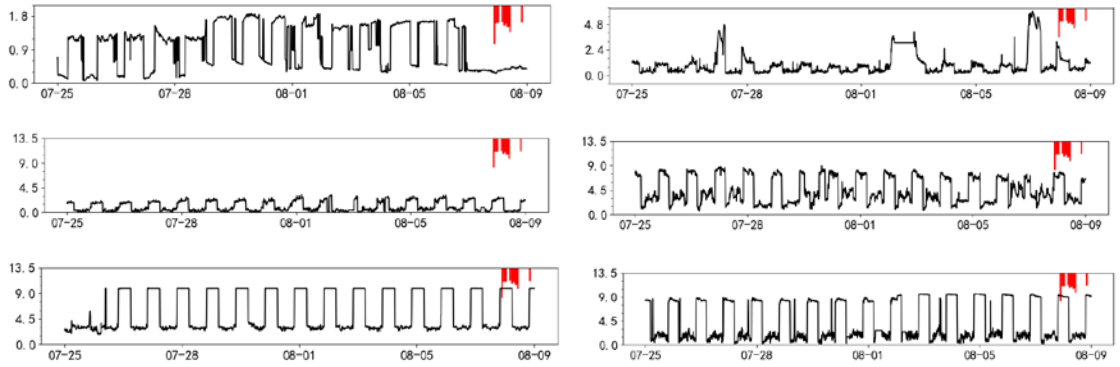


图 2.10　SRSS 波形图

进一步地，作者统计了九寨沟当地日升日落时间与九寨沟防震减灾局台站的低频电磁均值数据的 SRSS 波起落时间差，发现在地震发生前三天，当地日升日落时间与 SRSS 波起落时间明显减小，SRSS 波形起落时间与日升日落时间几乎一致，如图 2.11 所示。

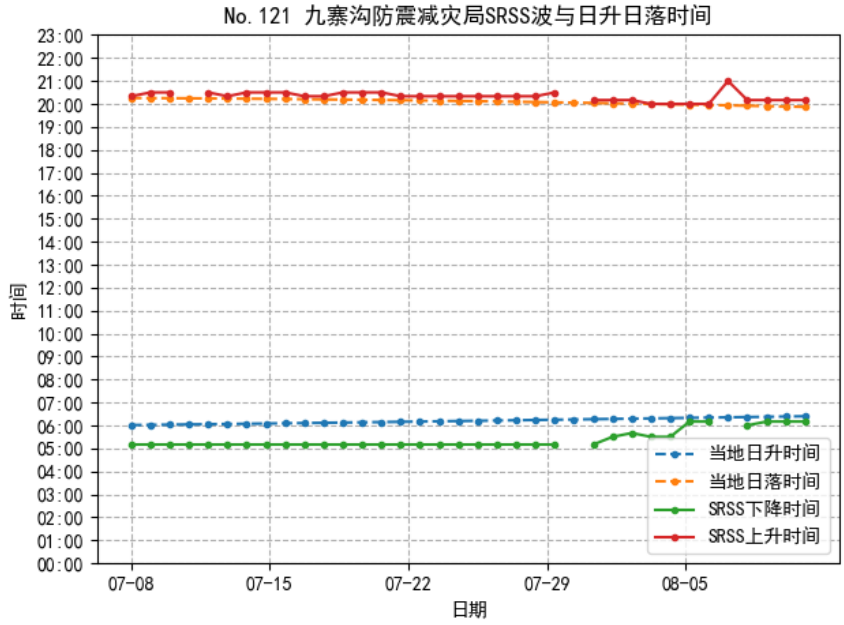



图 2.11　九寨沟防震减灾局 SRSS 波与日升日落时间图

另外，本文对于电磁均值数据进行了主成分分析（PCA），也取得了一些进展。2017 年 8 月 8 日九寨沟 7.0 级地震前后 AETA 电磁均值数据的主成分，在震前 PCA 值出现了明显异常，震后异常消失，如图 2.12 所示。

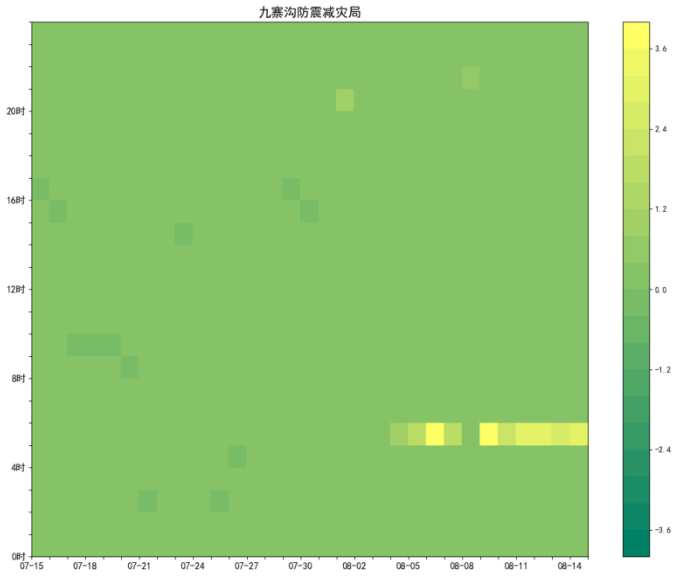



图 2.12　九寨沟防震减灾局 PCA 异常

### 2.4.2　地声数据震例分析

在地声数据的分析工作中，作者发现了一些与地震事件相关的信号，相比较于电磁数据，地声数据的前兆时间更短，没有体现地震孕育的过程，往往是地声数据发生较大波动后，台站周边会有一定可能性发生地震。这个特点在峨眉山防震减灾局（EMS）非常明显，在其他部分台站也有一定的映震效果，如图 2.13 所示。

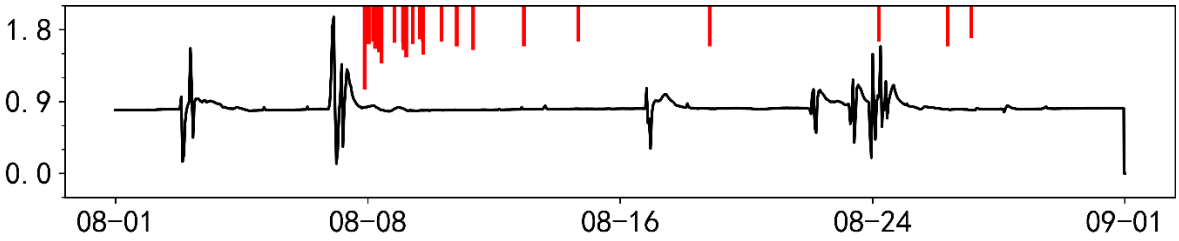



图 2.13　峨眉山防震减灾局地声数据异常

## 2.5　本章小结

本章介绍了 AETA 的系统架构、原始数据处理和分析情况以及 AETA 系统中的部分典型震例分析。首先，介绍了 AETA 由终端、电磁探头、地声探头以及云端中间件构成。其次，研究了电磁探头和地声探头采集的信号的物理意义、缺失数据处理方法以及原始数据的特征提取及分析。最后，分别对电磁数据和地声数据进行了部分震例分析，得到了 AETA 系统数据具有一定的映震效果，但是在数据处理和地震三要素的预测问题上，依然需要更先进的算法的支撑。

# 第三章　基于 AETA 数据的特征空间生成

上一章根据 AETA 原始数据从时域、频域两个方面提取了丰富的特征序列，本章是对这些特征序列进行描述，将序列的形态描述作为模型的输入特征。序列的描述使用时间序列数据挖掘方法，分割序列，计算波形之间的相似度，近似匹配序列，得到序列描述。进一步，本章详细叙述了样本打标的方法。本文将地震按照震级范围分为了 4 个级别，分别生成了 4 个特征数据集，每个数据集对应一种模型，每个模型负责对一个震级范围的地震预测。另外，特征序列的时间窗口作为地震发生的时间。

现有的时间序列数据挖掘问题大致可以分为 2 类，其一是可以清楚明确地知道每种时序对应的物理意义。譬如语音识别、人体姿态检测等问题，这种问题通常可以通过对时间序列数据标注，采用有监督机器学习方法来解决；其二是无法清楚明确地知道时序波形对应的物理意义，通常原因是该时间序列数据是一个多变量函数，造成时序波动或者异常的原因非常多。譬如地震数据，这种情况一般有两种可行的思路，第一种首先对时序数据进行降噪，将可能产生的噪声滤掉，从而可以将问题转化为第一类问题求解；第二种是将时序数据转化成结构化数据，利用结构化数据建模对问题进行求解。由于对数据进行降噪需要利用控制变量的方法尽可能地找出噪声，而对于地震事件，并没有成熟的开展实验的条件，所以本文采用第二种方法来生成所需的特征空间，求解问题。

## 3.1　基于时序数据挖掘的特征序列的描述方法

时间序列数据挖掘技术对于类似 AETA 这类长期监测系统非常重要，在这类系统中，信息的主要呈现形式便是时间序列。在 AETA 系统中，所有的信息都包含在电磁扰动探头和地声探头所记录的时间序列数据中，为了有效的将这些时间序列数据应用到模型中，本文需要一整套方法来对时间序列数据信息挖掘，并且通过时间序列的描述来呈递数据中潜在的有效信息。考虑到 AETA 系统数据的特点，本文将问题分成两类来考虑。第一类，对于单个时间序列数据的处理，这类问题主要考虑单个时间序列数据时间先后之间的联系以及时序数据内在的信息。第二类，对于多个台站的多个时间序列数据的处理，这类问题主要考虑多个台站之间的联动关系，找出台站之间的共同特征。

单个时间序列数据挖掘方法：

目前，对于单个时间序列数据挖掘方法主要有矩阵轮廓(Matrix Profile)、相似性匹配、字典学习、时间序列模式识别、时间序列分割等。矩阵轮廓是 Hoang Anh Dau 等

人提出的提取时序主题的方法，在 ECG、加速度传感器等序列中，用于提取关键模式，得到了不错的效果。例如，在心电信号中，将主要的心电波形与设备校准信号识别出来，在人体运动检测信号中，将俯卧撑、走路、跑步等信号识别出来。

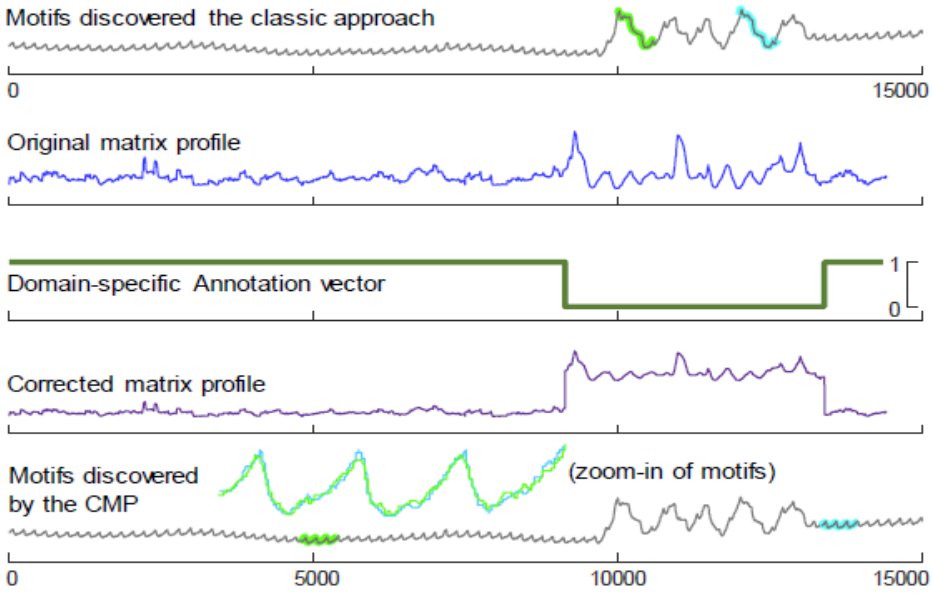



图 3.1　信号识别图

多台站时间序列数据挖掘方法：

对于同一个对象的多个传感器的时间序列数据挖掘方法主要有 TICC[68]、Tripoles[69]等。其中 TICC 是 David Hallac 等人在 2017 年提出的基于托普利兹逆矩阵的多个时序间相关性聚类的方法，将时序分割后，通过求解各类别信号的逆协方差矩阵来对信号进行分类。Tripoles 是 Saurabh Agrawal 等人提出的一种多种时序之间的新型关系，通常只会考虑两个时序间的相似度或者关联度，来对多个时间序列数据进行分析。但是 Tripoles 提供了一个新的思路，考虑 3 个时间序列之间的相关性，这样就可以在多个时序中找出相关性最高的 3 个，从这 3 个时序之间的固有关系中发现更多的信息。

针对 AETA 系统的数据特性，本文也设计了一系列的单一时序、多台站时序的数据挖掘方法，来呈递本系统中的时序特征。其中主要有：SRSS 波形的精准识别方法、AETA 时序的模态识别、AETA 时序的特征描述、多台站时序相关性分析。下面，对这四种方法展开讨论。

### 3.1.1　SRSS 波形识别方法

SRSS 波形是 AETA 系统电磁扰动均值数据中的一类特殊波形。在多个地区多个台站 AETA 的观测数据，均显示出近似日周期的变化，变化的形态包括类方波、双峰波、单峰波、以及不规则波动等形态，不管在地震多发地区如四川、云南，还是在地震较少的地区如河北、广东，近似日周期的波形都有出现，如图 3.2 至 3.7 所示。

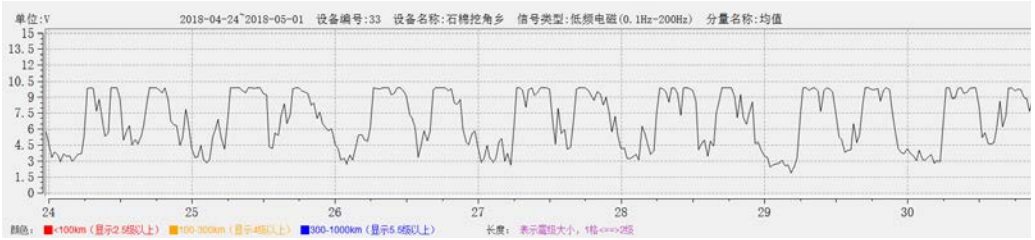



图 3.2　四川省石棉挖角乡的近似日周期波形：2018-04-24 至 2018-04-30

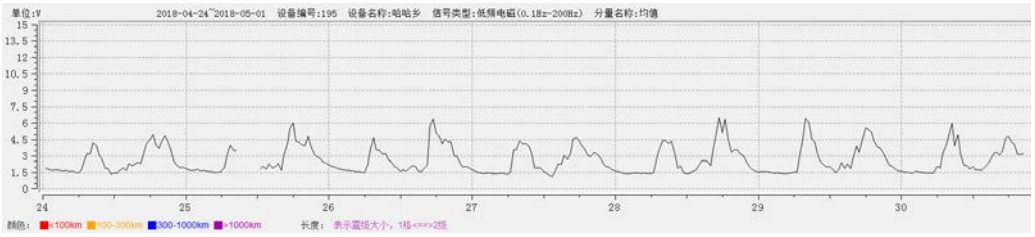



图 3.3　四川省冕宁哈哈乡的近似日周期的波形：2018-04-24 至 2018-04-30

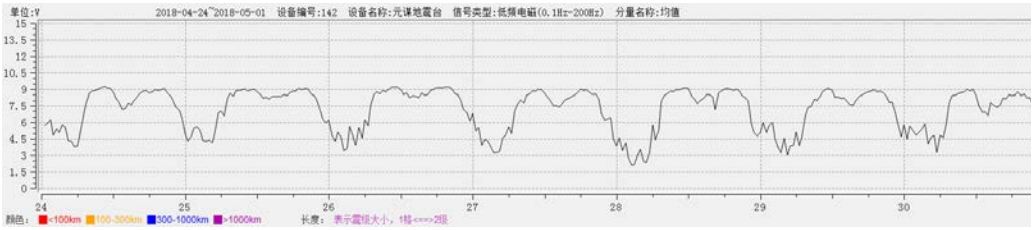



图 3.4　云南省元谋县的近似日周期波形：2018-04-24 至 2018-04-30

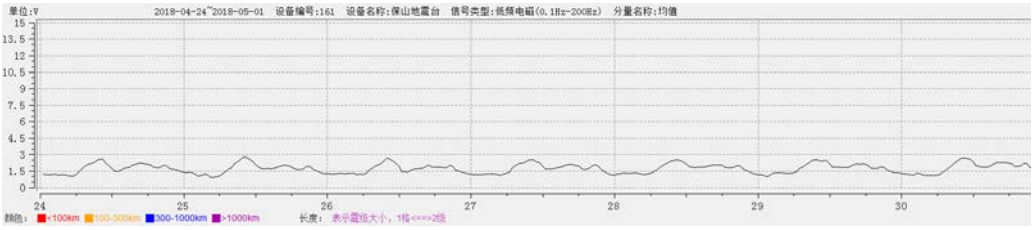



图 3.5　云南省保山市的近似日周期波形 2018-04-24 至 2018-04-30

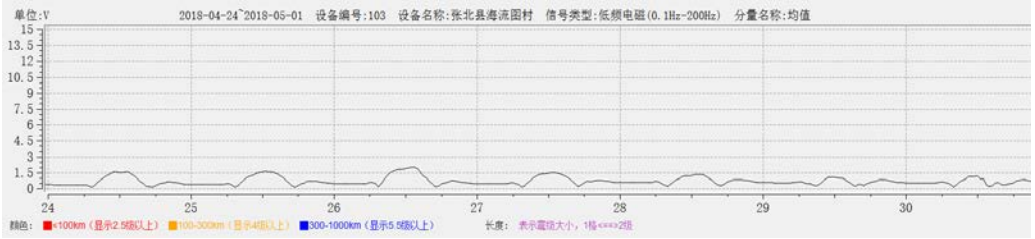



图 3.6　河北张北县的近似日周期波形：　2018-04-24 至 2018-04-30

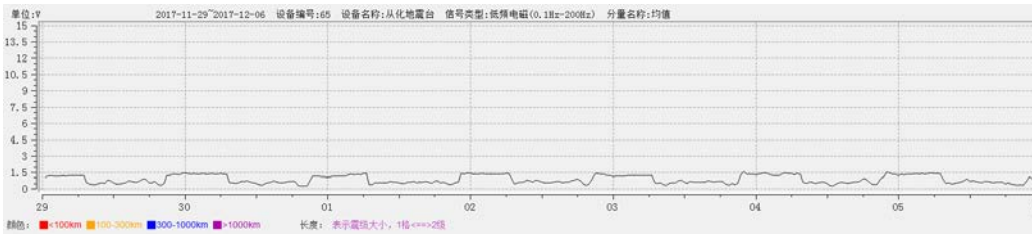


图 3.7　广东省从化的近似日周期波形：2017-11-29 至 2017-12-06

AETA SRSS 波

近似日周期的波形中，有一种与台站当地日升日落几乎同步变化的以天为周期的周期波形，日升时变低、日落时变高，课题组称之为“SRSS 波”。图 3.8 和 3.9 给出了九寨沟防震减灾局和唐山滦南的 SRSS 波形说明。

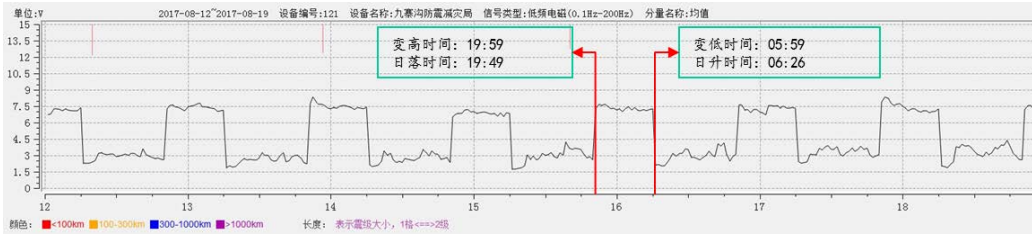



图 3.8　九寨沟防震减灾局的 SRSS 波：2017-08-12 至 2017-08-19

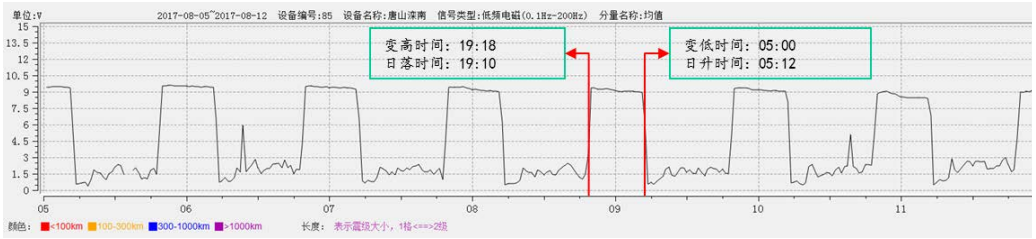



图 3.9　唐山滦南的 SRSS 波：2017-08-05 至 2017-08-12

根据第二章 2.4 节中的震例分析，本文发现了 SRSS 波形以及其变异波形对于地震事件有较强的指示作用，因此对于这类波形的识别和提取工作显得尤为重要。根据历史数据的整理和总结，本文得到了如下 SRSS 波形的数学定义以及识别方法。

SRSS 波形识别的数学定义

上升沿识别：

(1)一天的数据做归一化（0-1）Min-Max Normalization 处理，如式（3.1）所示。

$$x_i^* = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

其中，$x^*$ 为归一化后的数据，$x$ 为原始数据，$x_i$ 表示原始数据的第 i 个点。

(2)取台站当地日落时间前后 1 小时数据形成子序列 Ts，对于子序列 Ts，

(3)以 3 个点为滑窗，遍历 Ts，得到连续 3 个点平均斜率 K 的绝对值最大（Kmax）的时刻，如式（3.2）、（3.3）所示，若此时 Kmax 大于阈值 u，则判定为上升沿，如式（3.4）所示。

$$K_i = \frac{\left|x^*_{i+1} - x^*_i\right| + \left|x^*_{i+2} - x^*_{i+1}\right|}{2} (t_{sunset} - 20 \le i \le t_{sunset} + 20) \tag{3.2}$$

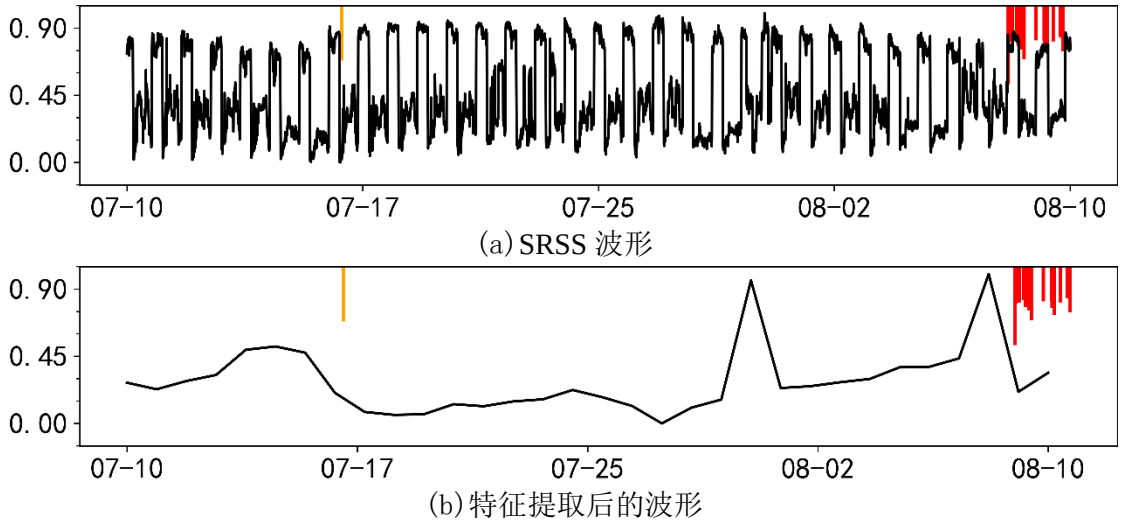
$$K_{\max} = \max(K) \tag{3.3}$$

$$if\left(K_{\max} > u\right), \arg K_{\max} \text{ 为上升沿/下降沿时刻} \tag{3.4}$$

下降沿识别：一天的数据做归一化（0-1）处理，如式（3.1）所示，取台站当地日升时间前后 1 小时数据形成子序列 Ts，对于子序列 Ts，以 3 个点为滑窗，遍历 Ts，得到连续 3 个点平均斜率的绝对值最大（Kmax）的时刻，如式（3.2）、（3.3）所示，若此时 Kmax 大于阈值 u，则判定为下降沿，如式（3.4）所示。

同步识别： 如果 TR(转换到达 50%（高值-低值）/2 时候的时间) 与台站当地日升日落时间差值在一个小时以内，则判定为与日升日落同步；每天一个日升判定，一个日落判定，取交集，同时符合才可以。

高值保持识别：对于高值区间的数值做归一化（-1 到 1）处理，使得均值为 0,如果标准差小于等于 1 则判定为符合高值保持。




(a) SRSS 波形




(b) 特征提取后的波形

图 3.10　SRSS 波形特征提取

### 3.1.2　AETA 时序模态识别

1) 序列分割

要对时间序列数据信息进行挖掘，就需要对不同粒度下的时序进行统计，从而得到不同时间尺度下的时序特征。根据第二章 2.4 节 AETA 观测数据震例分析中所述，本文从天周期、周周期、月周期为尺度，对时间序列数据进行分割，如图 3.11 所示。

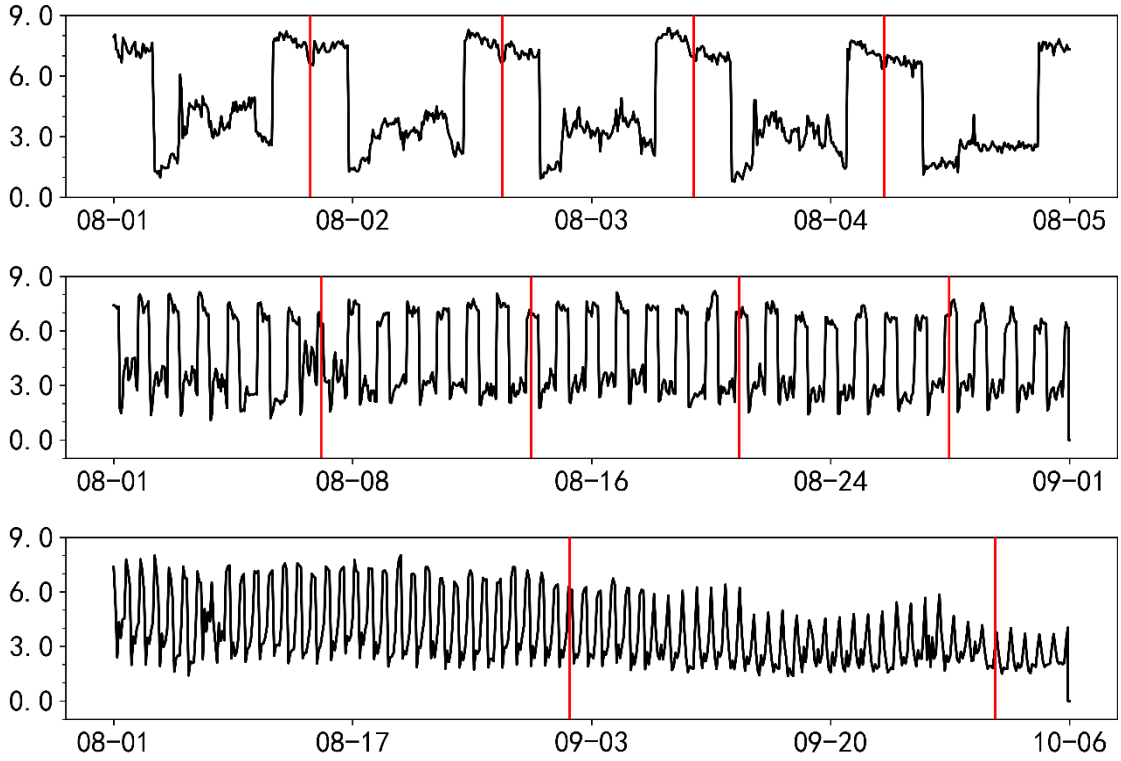



图 3.11　上、中、下分别为：天周期，周周期，月周期数据分割

2) 基础波形提取

首先，本文对 AETA 系统中所有台站在 2017-06-01 和 2017-09-01 期间按天分割的序列进行查看，从中定义了 7 种重要波形的形态，以这 7 类波形的归一化后的均值作为基础波形。具体分类情况如下图所示。

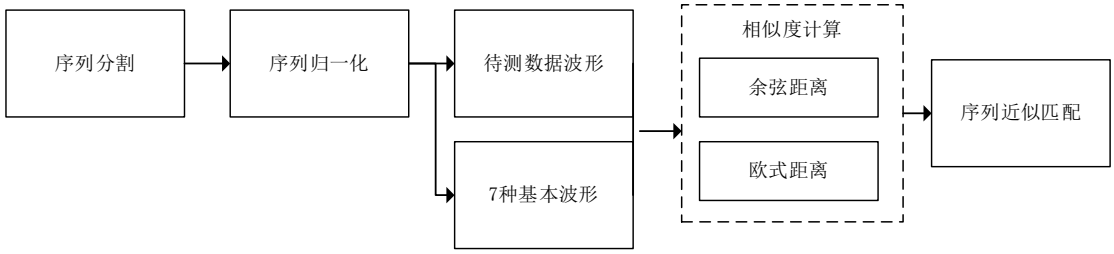



图 3.12　波形模式识别流程

3)相似度计算

提取到基础波形后，本文利用欧氏距离和余弦距离来衡量波形间的相似度，将每天的波形看作一个 24*1 维列向量，令基础波形为 x1，待测波形为 x2，则衡量欧式相似度的公式如下：

$$d = \sqrt{\sum_{i=1}^{N}(x_{1i} - x_{2i})^2} \tag{3.5}$$

衡量余弦相似度的公式如下：

$$\cos\theta = \frac{\sum_{i=1}^{n} x_{1i} \times x_{2i}}{\sqrt{\sum_{i=1}^{n} (x_{1i})^2} \times \sqrt{\sum_{i=1}^{n} (x_{2i})^2}} = \frac{X_1^T \cdot X_2}{\|X_1\| \times \|X_2\|} \tag{3.6}$$

4)序列近似匹配

基于上述原理，本文设计了如下算法对 AETA 系统中的均值特征数据进行波形识别与分类：

表 3.1　波形模式识别算法

**算法 3-1 波形模式识别算法**

**Input** Time series: base wave BW, objective wave OB,.

**Output** Similarity of two time-series.

1: initialize $\tau \leftarrow \emptyset$

2: **for** each $OB_i \in$ stations **do**

3:　**if** $\eta$==0 **then**

4:　　$\tau \leftarrow \tau$　calculate similarity with cosine distance

5:　**else if** $\eta$==1 **then**

6:　　$\tau \leftarrow \tau$　calculate similarity with Euclidean distance

7:　**end if**

8: **end for**

9: **return** Similarity$\leftarrow \tau$

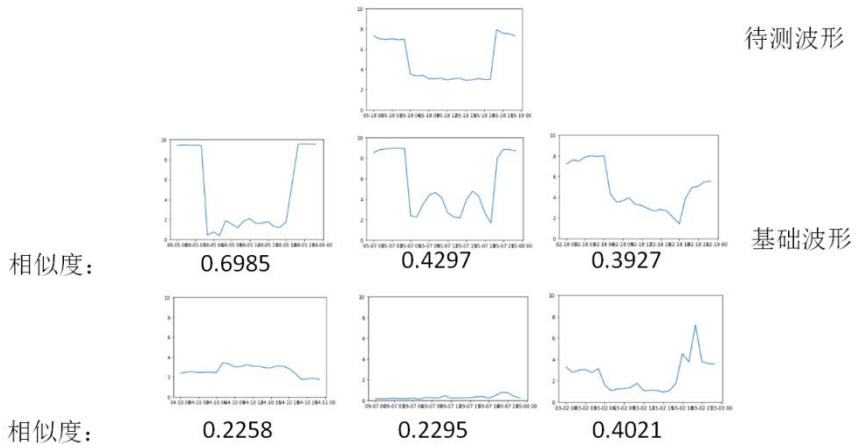



图 3.13　波形相似度分析

图 3.13 即为波形相似度分析，其中第一行为待测波形，第二第三行为已知的基础波形，下方的数字即为相似度。

5)序列描述

通过序列匹配结果形成对波形的描述。

通过上述算法，本文可以对于任意一天的数据进行宏观上的波形描述，形成如下类别特征：

表 3.2　波形类别特征

| 类型 | 类别特征 |
|---|---|
| SRSS | 0 |
| right_high | 1 |
| middle_2_peak | 2 |
| left_high | 3 |
| reverse_square_wave | 4 |
| flat_wave | 5 |
| middle_long_peak | 6 |
| others | 1 |

### 3.1.3　AETA 时序特征描述

根据 3.1.2 小节中所述的波形模式识别方法，可以对 AETA 系统电磁数据按天进行描述。所得到的结果即为台站 A 在当天的波形模式形成的序列。

### 3.1.4　多台站时序相关性分析

AETA 系统具有多个台站监测同一物理量的特点，为了更好的提取多个台站的联动信息，矩阵分析是一种可行的做法。常见的矩阵分析方法有奇异值分解（Singular Value Decomposition，SVD）、QR 分解、主成分分析（Principle Component Analysis，PCA）等。人类对于矩阵的感性认知和理解远远不如单一的数字深刻和敏感，因此这些矩阵分析方法都有一个共同点——简化矩阵，在损失信息尽可能少的情况下，用更低维的矩阵来描述原有的高维复杂矩阵。AETA 系统的特征数据每三分钟一个点，每天一个台站有 480 个数据，一共近 300 个台站，这样组成的矩阵是一个非常高维的矩阵，很难人工理解这些数据，在这样的情况下，利用 PCA、SVD 等方法先对矩阵进行降维，显得十分必要。下面，本文对 PCA、SVD 等方法进行介绍。

奇异值分解(SVD)是一种线性代数中的矩阵分解方法。在机器学习领域，SVD 方法应用非常广泛，不论是在多元时间序列、图像处理还是在推荐系统中都有着非常广泛的应用。奇异值分解最重要的作用就是抽取矩阵中重要的特征，相比于特征值分解只能对方阵进行，奇异值分解可以对任意形状的矩阵进行分解，扩大了自由度。由于 AETA 系统的时间序列采样频率较高，数据点比较多，为了减小数据量、降低特征维

度，应用 SVD 方法对多个台站数据做奇异值分解是有一定必要性的。而且奇异值分解不但起到了重要特征提取的作用，还起到了降低数据噪声的作用。

奇异值分解的原理如下：

假设 A 是一个 N * M 的矩阵，那么得到的 U 是一个 N * N 的方阵（里面的向量是正交的，U 里面的向量称为左奇异向量），Λ 是一个 N * M 的矩阵（除了对角线的元素都是 0，对角线上的元素称为奇异值），$V^T$(V 的转置)是一个 N * N 的矩阵，里面的向量也是正交的，V 里面的向量称为右奇异向量）。

$$A = U \cdot \Lambda \cdot V^T$$

通常前 10%的奇异值的累积和就占据了所有奇异值累积和的 90%以上，因此，奇异值可以很好的反映高维矩阵的信息。

主成分分析

主成分分析法（Principle Component Analysis）是一种矩阵降维方法。与 SVD 不同的是，PCA 更加关注信号的方差，它认为方差大的方向是包含信息量最大的方向，因此以方差最大的方向作为第一主成分方向，形成一组正交基，将原矩阵线性变换到这组正交基下表示，这样就起到了降维和特征提取的作用。PCA 算法流程如下：

输入：n 维样本集 $D = (x^{(1)}, x^{(2)}, ..., x^{(m)})$，要降维到 n’。

输出：降维后的样本集 $D'$

1）对所有样本进行中心化：$x^{(i)} = x^{(i)} - \frac{1}{m}\sum_{j=1}^{m} x^{(j)}$

2）计算样本的协方差矩阵 $X \cdot X^T$

3）对矩阵 $X \cdot X^T$ 进行特征值分解

4）取出最大的 n’个特征值对应的特征向量 $(\omega_1, \omega_2, ..., \omega_{n'})$，将所有特征向量标准化后，组成特征向量矩阵 W。

5）将样本集中的每一个样本线性变换到矩阵 W 组成的空间中，得到 $z^{(i)} = W^T \cdot x^{(i)}$。

6）得到输出样本集 D’。

在本文工作中，结合 AETA 系统的数据，为 PCA 主要关注在样本之间的方差上，而之前的工作表明这种信号在同一台站不同时间的数据分析中有一定效果，但是在多台站数据分析中，效果并不是很好，因此，多台站的时序相关性分析主要利用 SVD 方法来进行。下图所示为本文提出的基于 AETA 系统数据的多台站特征提取方法流程图以及具体做法。

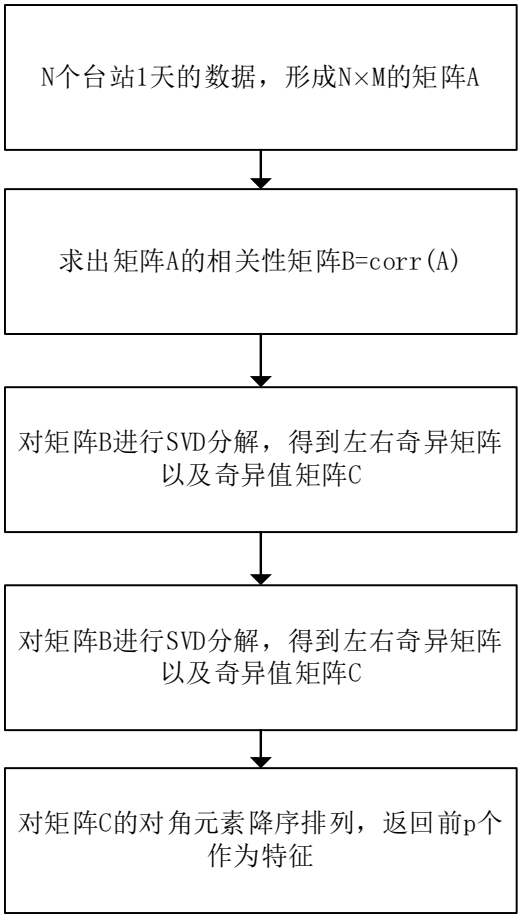



图 3.14　基于 AETA 系统数据的多台站特征提取方法流程图

表 3.3　多台站特征提取方法

**算法 3-2** 多台站特征提取方法

**输入：**研究地理范围、研究时间范围、p

**输出：**特征矩阵 D

1.在研究地理范围内找到所有 N 个台站

2.按天获取这 N 个台站的数据，得到 N*M 的矩阵 A（M 为 1 个台站 1 天的数据数）

3.计算得到矩阵 A 的相关系数矩阵 B=corr(A)

4.对 B 进行 SVD 分解，得到左奇异矩阵 u，右奇异矩阵 v，奇异值矩阵 C

5.C 的对角元素降序排列，返回前 p 个作为当天的特征向量，追加到 D 中

6.若下一天在研究时间范围中，则返回 2. 否则返回 7.

7.返回特征矩阵 D

图 3.15 中展示了本文在研究区域内对 38 个台站数据进行 SVD 特征提取后的结果，可以看到，地震发生前后，该区域内 SVD 特征值也较大，而无地震时，区域内 SVD 值相对较小。

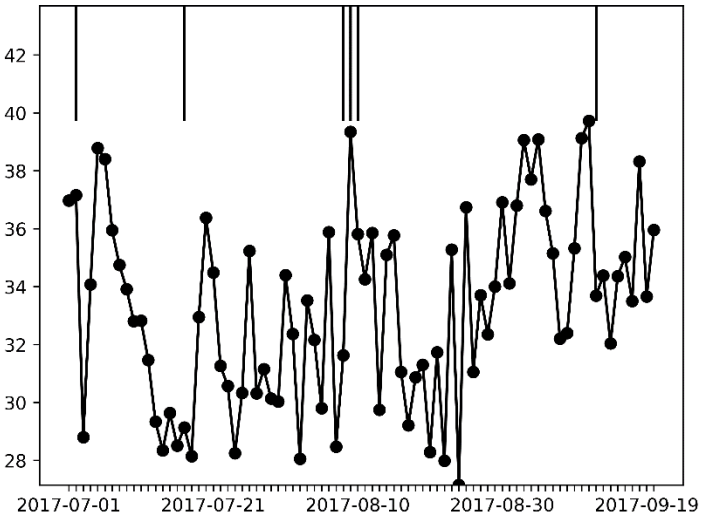



图 3.15　SVD 特征序列

## 3.2　基于地震事件密度的特征提取

如第一章和第二种中的内容所述，地震事件成因往往十分复杂，到目前为止，在学术界还没有一个定论，这也给本文的工作带来了巨大的困难。形成原因不同的地震，往往也伴随不同的前兆信号，AETA 系统从开始布设到现在共经历了 2 年时间，在这期间，国内地震最频发的川滇地区一共发生了 32 次 4 级以上地震，261 次 2 级以上地震。本文研究发现，不同时间段的地震频数并不相同，作者对 2017 年 3 月 1 日到 2019 年 2 月 28 日之间， [97° E, 109° E, 25° N, 35° N]范围内的地震事件根据震级进行了统计。统计发现，2017 年 8 月份地震次数最多，主要由于九寨沟 7.0 级地震及其后续的一系列密集的余震。此外，2017 年 4 月以及 2018 年 4 月都没有大于 4 级地震发生。而在研究范围内，平均每个月会有 1.6 次 4 级以上地震发生，12.3 次 2 级以上地震发生。从时间规律角度来看，本文的统计并没有发现明显与月份相关联的地震频次或者地震震级规律，如表 3.4 所示。

表 3.4　地震事件表

| | 地震频数 | | 震级叠加 | | AETA 电磁 | AETA 地声 |
|---|---|---|---|---|---|---|
| | >4 级 | >2 级 | >4 级 | >2 级 | | |
| Mar-17 | 4 | 18.6 | 10 | 37.3 | 22.3 | 2.05 |
| Apr-17 | 0 | 0 | 6 | 20.2 | 19 | 1.06 |

续表 3.4　地震事件表

| | | | | | | |
|---|---|---|---|---|---|---|
| May-17 | 2 | 9.3 | 11 | 38 | 17 | 0.96 |
| Jun-17 | 0 | 0 | 7 | 22.6 | 19 | 0.82 |
| Jul-17 | 2 | 9 | 9 | 29.2 | 20 | 0.86 |
| Aug-17 | 4 | 20.2 | 46 | 155.2 | 22 | 1.18 |
| Sep-17 | 2 | 9.8 | 18 | 59.7 | 23.3 | 0.92 |
| Oct-17 | 2 | 9.6 | 14 | 45.5 | 21.4 | 0.95 |
| Nov-17 | 3 | 13.7 | 9 | 32.5 | 21.9 | 1.07 |
| Dec-17 | 0 | 0 | 7 | 20.9 | 22.3 | 1.55 |
| Jan-18 | 0 | 0 | 4 | 11.9 | 21.8 | 1.5 |
| Feb-18 | 3 | 12.6 | 10 | 34.4 | 19.6 | 1.74 |
| Mar-18 | 0 | 0 | 15 | 45.3 | 16.8 | 2.06 |
| Apr-18 | 0 | 0 | 10 | 28.1 | 17.9 | 1.82 |
| May-18 | 2 | 8.6 | 19 | 60 | 23 | 1.23 |
| Jun-18 | 1 | 4 | 6 | 19.7 | 23.9 | 1.51 |
| Jul-18 | 1 | 4.2 | 7 | 24 | 26 | 1.61 |
| Aug-18 | 1 | 4.4 | 16 | 51.7 | 25.4 | 1.44 |
| Sep-18 | 1 | 5.3 | 5 | 18.3 | 29.8 | 1.31 |
| Oct-18 | 2 | 9.6 | 10 | 33.6 | 25 | 1.79 |
| Nov-18 | 0 | 0 | 7 | 21 | 25.7 | 2.07 |
| Dec-18 | 3 | 15 | 14 | 49.2 | 25.44 | 1.89 |
| Jan-19 | 1 | 5.3 | 20 | 66.6 | 22 | 1.94 |
| Feb-19 | 4 | 18 | 15 | 53 | 24.14 | 1.91 |

表 3.4 分别统计了研究范围内的大于 4 级地震频数、大于 2 级地震频数、大于 4 级地震震级叠加、大于 2 级地震震级叠加以及 AETA 系统相应的电磁扰动数据和地声数据。其中，震级叠加是指，将一个月内发生的符合条件（>4 级或者>2 级）的地震震级累积求和，以表征该月通过地震释放的总能量。AETA 电磁数据的统计是将研究时间范围和地理范围内的所有台站的低频电磁均值数据进行聚合，按月计算均值，从而将该区域内一个月的低频电磁均值数据聚合为 1 个点，表征该月内该区域内的电磁数据能量。同理，作者对 AETA 低频地声均值也做相同的处理，表征该月内该区域内的

地声数据能量。图 3.16 中展示了地震频数和震级叠加在不同时间段均有不同的现象。图 3.17 展示了在这个过程中 AETA 数据与地震频数和地震叠加之间的相关性。

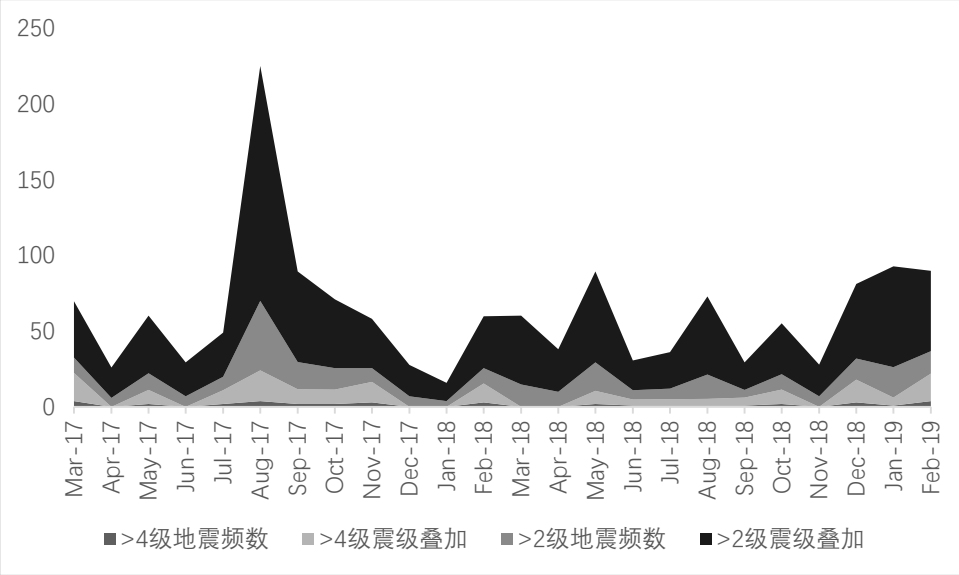



图 3.16　地震频数与震级叠加图

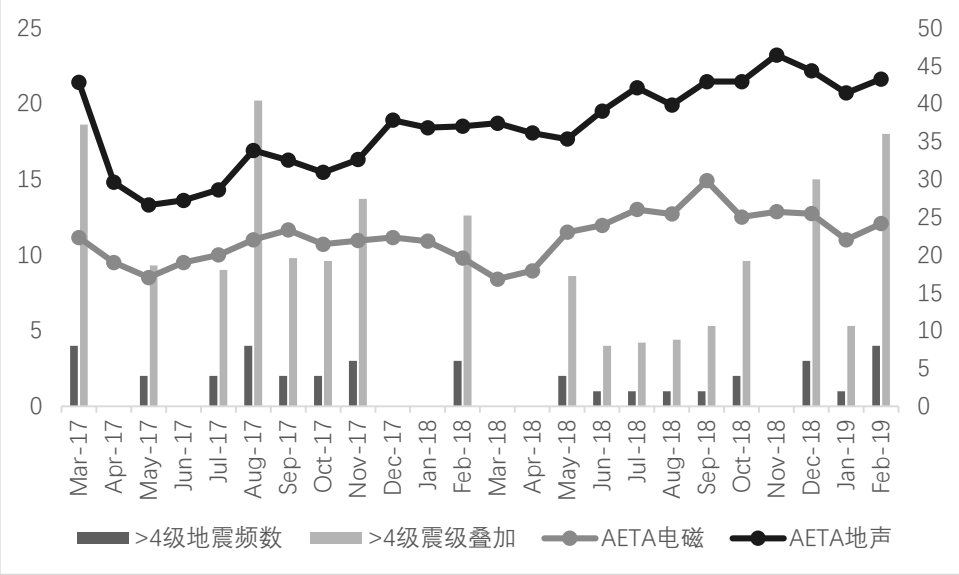



图 3.17　AETA 数据与地震频数和地震叠加

如图 3.17 所示，AETA 系统的电磁、地声数据按月聚合为一个点后，与地震的频率和震级叠加有一定的相关性。在 2017 年 5 月-8 月，4 级以上的地震频数趋势与 AETA 电磁、AETA 地声数据的趋势保持一致。2018 年 2 月-5 月，4 级以上的地震趋势与 AETA

电磁数据的趋势基本一致。总体上来看，AETA 系统表征的能量与地震事件表征的能量，在趋势上，有一定相似性。

另外，经过作者长期对该区域地震事件的关注和研究，发现区域内有 4 级以上地震时，往往在接下来一段时间连续发生 4 级左右的地震，上图中的几个尖峰正是由于这个原因产生的。所以说知道前一次地震的信息，对于本文判断下一次地震信息是有帮助的，在统计学上，可以认为这是一种有效的先验信息。

因此，为了体现不同时间段内的地震频次对下一次地震预测结果的影响，本文将预测当天的前 d 天的地震频次以及这段时间内 AETA 系统的电磁、地声数据按月份的聚合值作为一组特征。

## 3.3 长短周期特征提取

在本章的前两节中，介绍了多种特征提取的方法，其中第一节主要介绍了以天为粒度的特征提取，第二节主要关注的是月为粒度的特征提取。然而，在实际数据观测中，背景值对数据，尤其是对较长时间粒度的数据影响较大，这种背景值与地震事件的关联性不大，体现的是地球内在的长周期运动结果。一般需要长达数年的观测才能找到其背后的规律，而本系统目前所积累的数据量还不足以很好的处理这样的背景值。为了提高数据质量，长短周期特征提取是一种可行的办法。

STA-LTA 算法在地震监测领域常被用于 P 波到时的震相提取[74]。一般来说，地震发生时产生的震动波与背景噪声在时域和频域上特征差别明显，而地震又是一个瞬发过程，因此该算法使用一个较小的滑动时间窗口来检测地震波信号，用一个较大的滑动时间窗口来检测背景噪声信号，最后通过小的时间窗口产生的特征与大的时间窗口产生的特征的比值来判断是否为震相到达时刻。采取这样策略的原因是，背景数据是一个随机序列，有时会在一个时间窗口中发生剧烈波动，这时候如果不考虑背景数据情况，很容易将这种情况误判为震相到时。当本文在 AETA 系统数据中提取一些长周期特征时，同样存在这样的情况，即背景值的波动会给本文提取的特征质量带来很大的影响。为了消除背景值给后续研究带来的不利影响，本文在特征提取时，采用 STA-LTA 算法。令 $A_i$ 为待研究序列，$N_{sta}$ 为短时窗口，$N_{lta}$ 为长时窗口，那么 i 时刻短时窗口的平均值 $STA_i$ 为：

$$STA_i = \frac{1}{N_{sta}} \sum_{i=1}^{N_{sta}} |A_i| \tag{3.7}$$

而 i 时刻的长时窗口的平均值 $LTA_i$ 为：

$$LTA_i = \frac{1}{N_{lta}} \sum_{i=1}^{N_{lta}} |A_i| \tag{3.8}$$

则，i 时刻短时窗口与长时窗口的平均值之比为：

$$\frac{STA_i}{LTA_i}=\frac{N_{lta}\sum_{i=1}^{N_{sta}}|A_i|}{N_{sta}\sum_{i=1}^{N_{lta}}|A_i|} \tag{3.9}$$

式(3.7)、(3.8)、(3.9)中的$|A_i|$可以替换为任何想要提取的特征，这里仅仅是以序列的绝对值（表征能量）这个特征值。这样，可以通过这种方法来有效消除长期背景值漂移等问题的影响。

## 3.4　地震三要素范围的选取

地震三要素是指地震发生时间、震级、震中位置，是地震预测核心的三个要素。在第二章中的 AETA 震例分析中可以看到，由于不同地区的地质结构和地理条件不同，不同的地震对应的前兆信号模式不尽相同，而地质结构往往非常复杂，涉及到的变量极多，短时间内难以出现两个发震机理和震中条件相似的地震，这就导致了地震三要素很难准确预测。因此，本文中对于三要素的研究均在一个可以接受的范围内进行。最后，根据预测效果选择最优的三要素范围。

### 3.4.1　时间窗口的选取

在地震预测领域，短临地震预测一般是指震前几小时到几十天这个时间范围的地震预测。这个时间窗口相比较于中长期地震预测确实更有科学意义和社会价值，但是不可否认，这个窗口从防震减灾角度来说，依然显得太大了。假设有研究者预报了未来 50 天某地区会发生破坏性地震，那么该地区需要在未来 50 天启动应急预案，严重影响区域内所有居民、企业、政府的生活和运转，才有可能避免接下来的破坏性地震带来的人员伤亡和财产损失，况且，预测都是基于概率的预测，难以做到 100%准确度，权衡下来，这个时间窗口依然难以在实际的防震减灾中起到积极的作用。

根据第二章所述的地震震例分析，可以看到 AETA 系统对于地震事件的前兆信号主要出现在 1-15 天这个范围。若最终选取的时间窗口为 n 天，则$n\in[1,15]$，目标是预测未来 n 天是否会有地震发生。令当前研究的时间为$T_0$，$T_{-n}$表示当前研究时间前 n 天，$T_{+n}$表示当前研究时间后 n 天。那么，如果在$T_0$时刻预测未来 n 天会发生地震，则当$[T_{+1},T_{+n}]$之间有地震发生，即认为预测正确，否则认为预测错误。如图 3.18 所示为预测正确和预测错误的情况示例。若在$T_0$时刻预测未来 n 天无地震，则当$[T_{+1},T_{+n}]$之间无地震发生，即认为预测正确，否则认为预测错误。

这个选取过程如式(3.10)-(3.12)所示：

$$\tilde{y}=F(x_{T_{-m}},x_{T_{-m+1}},...,x_{T_0}),\ \tilde{y}\in[-1,+1] \tag{3.10}$$

$$tl_i=g(T_i),\ tl_i\in[-1,+1] \tag{3.11}$$

$$y=\begin{cases}+1, & \max(tl_i)=1\\ -1, & \max(tl_i)=-1\end{cases} \tag{3.12}$$

其中，$\tilde{y}$ 为根据 $T_{-m}$ 至 $T_0$ 时刻历史数据预测未来 $T_1$ 到 $T_n$ 时刻是否有地震发生的结果。$tl_i$ 表示 $T_i$ 时刻是否有地震，有为+1，没有为-1。$y$ 表示实际中 $T_1$ 到 $T_n$ 时刻是否有地震发生。那么，当 $\tilde{y}=y$ 时，说明此次预测正确，$\tilde{y}\neq y$ 时，说明此次预测错误。

| 时间 | 是否地震 |
|---|---|
| $T_0$ | N |
| $T_{+1}$ | N |
| $T_{+2}$ | Y |
| … | … |
| $T_{+n}$ | N |

预测正确

| 时间 | 是否地震 |
|---|---|
| $T_0$ | N |
| $T_{+1}$ | N |
| $T_{+2}$ | N |
| … | … |
| $T_{+n}$ | N |

预测错误

图 3.18　预测正确和预测错误的情况示例

接下来讨论如何确定时间窗口 n 的值。如前文所述，不同的地震通常对应的前兆信号时间窗口也不同，由于地震成因问题至今在学术界也没有定论，所以难以从机理层面着手来处理这个问题。因此，本文从结果出发，利用预测准确率反过来确定时间窗口 n 的值，这里可以把 n 理解为地震预测模型中的一个超参数。在机器学习领域，常常利用交叉验证的思想来确定模型中的超参数取值，本文亦借鉴这种思想来确定时间窗口 n 的值。具体地，先将样本进行 5 折交叉划分，用训练集进行模型训练，记录在模型验证集上的预测准确率，然后在其他条件不变的前提下，仅改变 n 的取值。记录在验证集上准确率最高的 n 值作为最终的时间窗口。具体操作图 3.19 所示：

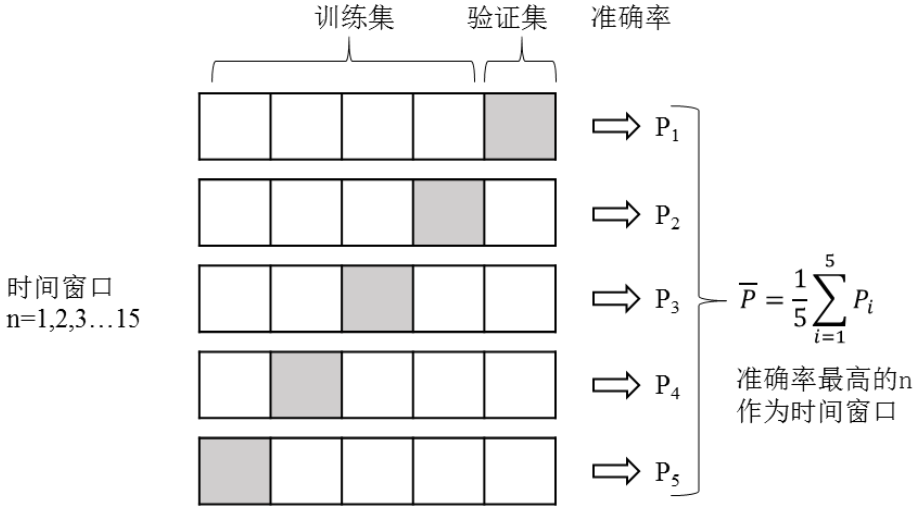



图 3.19　时间窗口

### 3.4.2　震级范围的选取

一般来说，将小于 1 级的地震称为超微震，1 级到 3 级的地震称为微震，3 级到 4.5 级的地震称为有感地震，4.5 级到 6 级的地震称为中强震，6 级到 7 级的地震称为强震，7 级以上的地震称为大地震[73]。根据中国地震台网统计，2018 年 3 月 1 日-2019 年 3 月 1 日期间，我国共发生 3 级以上地震 554 次，其中有感地震 526 次，中强震 26 次，大地震 0 次。近 5 年来，平均 27 天发生一次 5 级地震，116 天发生一次 6 级地震。图 3.20 统计了国内 5 级以上地震的分布情况，可以看到主要地点集中在川滇地区、西藏新疆以及台湾地区。

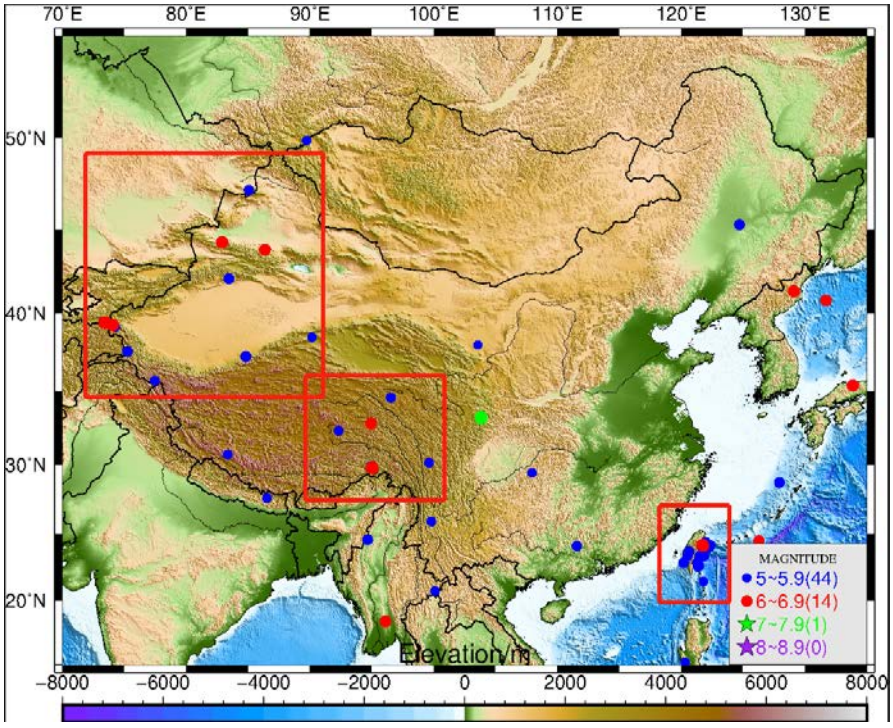



图 3.20　中国 5 级以上地震的分布情况

通过统计结果看到，有感地震次数远大于中强震和大地震次数。这样一来，就需要对地震事件按震级大小进行筛选，否则，少量的中强震和大地震会被大量的有感地震“淹没”，在后续的震级预测模型中难以被准确预测到。此外，考虑到不同震级的地震影响的范围不同，前兆表现也不相同，因此对震级进行离散化，有助于模型的精确度提升。

为了尽可能避免 AETA 台站密度较低的区域给研究带来的困难，本文选取 AETA 安装台站较密集的区域[97°E, 109°E, 25°N, 35°N]（如图 3.20 所示）作为研究区域。选取 2017 年 7 月 1 日-2019 年 3 月 1 日作为研究时间范围（2017 年 7 月 1 日该区域安装的第一批 38 套设备开始稳定运行）。在上述研究范围内，根据震级将地震分为有感

地震、中强震、强震、大地震三种，并分别统计了区域内共 609 天的地震情况，如表 3.5 所示。

表 3.5　地震情况表

| 震级 | 有地震天数 | 无地震天数 |
|---|---|---|
| 有感地震（3-4.5 级） | 172 | 437 |
| 中强地震（4.5-6 级） | 13 | 596 |
| 强地震（6-7 级） | 0 | 609 |
| 大地震（>7 级） | 1 | 608 |

可以看到，这样切分后，AETA 系统关注的中强震以上的地震只用 14 个，而有感地震占据了所有地震天数的 92%。中强震以上的震例不足够，因此，减少震级的分桶个数，并且将 4.5 级下调至 4 级，本文将地震按震级分为 0-4 级，4-7 级，7 级以上这三种。这三种地震在本文研究范围内的统计情况如表 3.6 所示。

表 3.6　地震统计

| 震级 | 有地震天数 | 无地震天数 |
|---|---|---|
| 3-4 | 153 | 456 |
| 4-7 | 32 | 577 |
| >7 | 1 | 608 |

这样，区域内中强震提升了 1 倍多，增加到了 32 个，正样本数量显著增多。在第五章中的模型建立中，本文就以这种切分方式来对样本打震级标签。

### 3.4.3　震中范围选取

AETA 设备在全国多个省、地区都有布设，但是由于多方面原因，不同省、地区的布设密度各不相同，部分发震区域布设密集，而部分发震区域布设量非常少，甚至没有布设，这为模型建立带来了困难。预测模型的输入信息量直接决定了预测效果，而地区间不同的布设密度代表了收集信息量的不同，这样建立的模型泛化能力必然不好。为解决密度不均带来的模型泛化能力问题，本文挑选了布设密度较为密集的区域作为研究对象，并且将这个区域按经纬度分割成了约为 111km*97km 的网格区域进行研究。如图 3.21 所示。

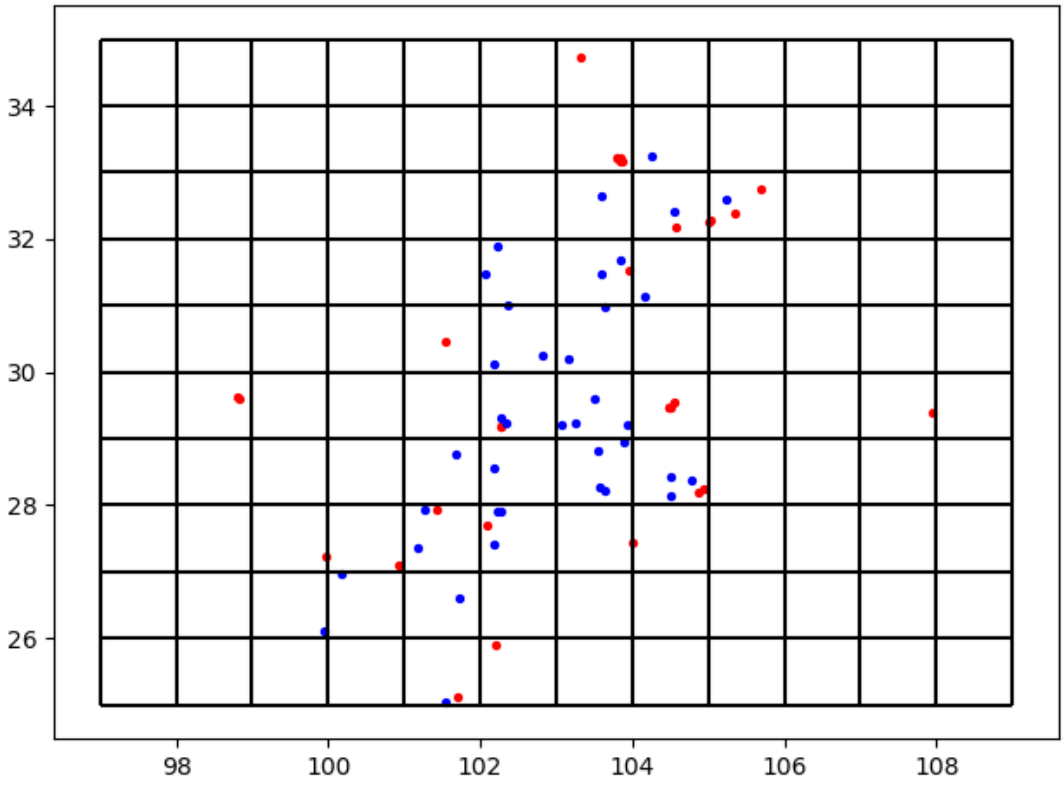



图 3.21　所需区域图

## 3.5　特征空间列表

表 3.7 给出了基于单台站时间序列数据挖掘、周期特征提取、多台站时序数据 SVD 特征提取、区域地震密度统计等方法得到的 32 个特征列表。

表 3.7　特征空间列表

| 特征 | 类型 | 含义 |
|---|---|---|
| Mag_SRSS_num | Int | 电磁 SRSS 数量 |
| Mag_SRSS_acc_max | Int | 电磁 SRSS 最大天数 |
| Mag_SRSS_acc_min | Int | 电磁 SRSS 最小天数 |
| Mag_SRSS_acc_mean | Float | 电磁 SRSS 平均天数 |
| Mag_category_0 | Category | 电磁类别为 0 的天数 |
| Mag_category_1 | Category | 电磁类别为 1 的天数 |
| Mag_category_2 | Category | 电磁类别为 2 的天数 |
| Mag_category_3 | Category | 电磁类别为 3 的天数 |

续表 3.7　特征空间列表

| | | |
|---|---|---|
| Mag_category_4 | Category | 电磁类别为 4 的天数 |
| Mag_category_5 | Category | 电磁类别为 5 的天数 |
| Mag_category_6 | Category | 电磁类别为 6 的天数 |
| Mag_category_7 | Category | 电磁类别为 7 的天数 |
| Sound_category_0 | Category | 地声类别为 0 的天数 |
| Sound_category_1 | Category | 地声类别为 1 的天数 |
| Sound_category_2 | Category | 地声类别为 2 的天数 |
| Mag_SVD_max | Float | 区域多台站电磁 SVD 最大值 |
| Mag_SVD_min | Float | 区域多台站电磁 SVD 最小值 |
| Mag_SVD_mean | Float | 区域多台站电磁 SVD 均值 |
| Mag_SVD_var | Float | 区域多台站电磁 SVD 方差 |
| Sound_SVD_max | Float | 区域多台站地声 SVD 最大值 |
| Sound_SVD_min | Float | 区域多台站地声 SVD 最小值 |
| Sound_SVD_mean | Float | 区域多台站地声 SVD 均值 |
| Sound_SVD_var | Float | 区域多台站地声 SVD 方差 |
| Mag_STA_LTA_count | Int | 电磁长短周期异常天数 |
| Mag_STA_LTA | Float | 电磁长短周期异常指数 |
| Sound_STA_LTA_count | Int | 地声长短周期异常天数 |
| Sound_STA_LTA | Float | 地声长短周期异常指数 |
| Eq_num | Int | 区域地震次数 |
| Eq_mag_max | Float | 区域地震最大震级 |
| Eq_mag_min | Float | 区域地震最小震级 |
| Eq_mag_mean | Float | 区域地震平均震级 |
| AETA_num | Int | 区域内 AETA 台站数量 |

## 3.6　样本不均衡问题及解决办法

根据上文所述的方法得到数据集后，发现对于 Ms>5.0 的地震事件存在较为明显的样本不均衡问题。对于该问题，本文采用一些过采样方法（例如 SMOTE[77]）对少数样本进行扩充。另外，由于不同地区的地震频率、地震前兆信号可能不同，后文中还将

进一步缩小研究区域，争取能够对一个更小区域的地震事件与 AETA 系统数据之间的关系进行更深入的研究。

## 3.7　本章小结

本章主要讨论了基于 AETA 系统数据的特征空间生成方式。AETA 系统的数据产出形式主要是时间序列。本章首先对于单个台站的一元时间序列数据进行特征提取，主要包括 SRSS 波形的识别，时序模态识别，波形描述等。其次对研究区域内的多个台站组成的多元时间序列进行特征提取。另外，本章中还统计了研究区域内的地震密度与震级之间的关系，并将地震频率作为特征加入特征集。最后，本章探讨了地震的时间、地点、震级三要素范围的选取以及样本标注的方法，并给出了特征空间的列表以及不均衡样本的解决方法。为后续模型的建立和进一步研究打下了基础。

# 第四章　基于关联分析方法的特征降维

上一章介绍了特征空间的生成，但是基于此方法生成的特征空间维度非常高，并且里面包含了噪声特征。再者，考虑到本文目前样本总量只有 523 个，所以本文需要一种降维方法，可以降低特征空间维度，去掉噪声。这样，既方便了后期模型的训练，又可以提升整个模型的可解释性。本章的降维采用 Apriori、FP-growth 等关联分析方法，这种方法通过计算集合对地震事件的支持度与可信度找出与地震事件相关的频繁项集，本文将频繁项集作为降维后的特征输入后续的模型。

## 4.1　关联分析方法

大数据中隐藏的关系可以通过挖掘事件之间的“关联规则”来探查。关联分析方法中，支持度与置信度是衡量事件集与结果之间是否构成关联的重要指标。支持度定义为某一特征出现的次数占所有特征出现次数的百分比；置信度定义为出现特定事件时，该特征也出现的概率。在本文中，本文试图找到特征序列的表现与地震事件的关联规则。

在挖掘样本中的频繁事项集合之前，往往需要先枚举整个样本空间内所有事项组合的结果，为了规范枚举的表达，学者们提出了一种格结构来表示对所有可能的项集的枚举结果，图 4.1 的格结构为事件集 $I=\{a,b,c,d,e\}$ 的所有事项枚举的结果。一般在实际场景中，k 值可能会非常大。而一个包含 k 个项目的样本集最多会产生 $2^k$ 种事项集的组合，这种指数增长关系使得需要探查的项集搜索空间变得非常大。

关联分析方法的常用场景是购物车分析，数据分析师通过统计不同消费者购物车中的物品，找出最常出现的组合，从而将合适的商品打包促销。这个最常出现的组合就是频繁项集。一般地，频繁项集的定义依赖于置信度和支持度两个指标。对于一个事件集 D，关联规则{A—>B}的置信度 r 表示在项集 A 出现在事件集 D 中的前提下项集 B 出现在事件集 D 中的概率。而关联规则{A—>B}的支持度则表示项集 A 和项集 B 同时存在于事件集 D 中的概率。当关联规则{A—>B}的支持度大于支持度阈值，置信度大于置信度阈值时，就认为该关联规则为一个频繁项集。其中支持度阈值和置信度阈值为认为设定的，典型值为 0.5。频繁项集发现的计算难度主要在于当项集空间大小随项数指数增长，从而造成搜索空间巨大，需要时间复杂度较高。

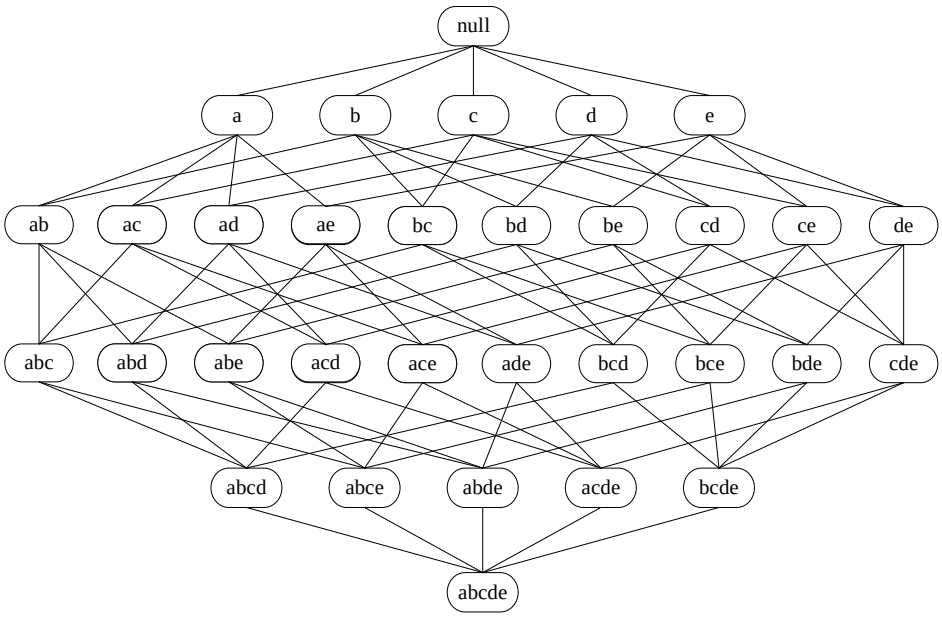



图 4.1　项集的格

一种简单的发现频繁事项集的思路是，枚举所有可能的组合来确定格结构中每个候选事项集的支持度，再根据支持度排序得到频繁项集。每次计算支持度都需要将每个候选项集与每个正(负)样本进行比较，如图 4.2 所示，如果候选项集包含在正样本中，则候选项集的支持度增加。例如，由于项集{电磁方波，地声异常}出现在事务 1,3,5 中，其支持度将增加 3。但是，这种方法的时间复杂度 O(NMw)非常大，其中 N 是样本数，$M = 2^k - 1$是候选项集数，而 w 是样本项集的最大长度，在本文中，将样本处理为等宽的结构化数据，因此样本中的 w 均相同。在 AETA 系统中，k 通常是一个非常大的值，这会导致搜索空间非常大。因此，必须使用算法帮助减小搜索空间。现有的频繁项集发现技术主要是 Apriori 算法和 FP-growth 算法，二者都可以通过一定策略来大幅降低搜索时间。

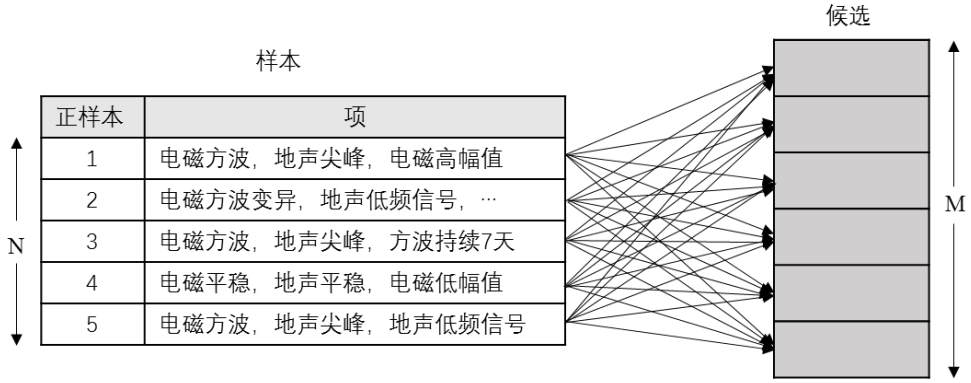



图 4.2　计算候选项集的支持度

### 4.1.1 Apriori

Apriori 算法是关联规则的挖掘算法通过计算项集支持度来对搜索空间进行先验剪枝，也是最早用于优化关联规则挖掘的时间复杂度的方法[78]。以算法 4-1 为例说明 Apriori 算法的先验剪枝技术。Apriori 算法的剪枝思想是：若{c, d, e}是频繁项，显而易见的是，所有包含事项集{c, d, e}的样本一定包含它的子集{c, d}，{c, e}，{d, e}。如此，若某项集是频繁的，则它的所有子集一定也是频繁的。否则它的所有超集也一定是非频繁的，如若发现一个关联规则是非频繁的，则整个包含该规则的超集的子图可以立即被剪枝。Apriori 算法依靠支持度阈值来判别某个候选事项集是频繁的还是非频繁的，若事项集支持度大于该阈值，则认为是频繁项集，若该项集支持度小于该阈值，则认为是非频繁项集。

表 4.1 Apriori 算法频繁事项集的产生

| 算法 4-1 Apriori 算法频繁事项集的产生 |
|---|
| 1: k = 1 |
| 2: $F_k$ = $\{i \mid i \in I \wedge \sigma(\{i\}) \geq N \times minsup\}$ / 发现所有的频繁 1-项集 |
| 3: while $F_k \neq \varnothing$: |
| 4: k = k + 1 |
| 5: $C_k$ <— 产生 $F_{k-1}$ 的候选项集 |
| 6: for 每个样本 $t \in T$ do |
| 7: $C_t$ <—识别出属于样本 t 的所有候选项集 |
| 8: for 每个候选项集 $c \in C_k$ do |
| 9: $\sigma(c) \leftarrow \sigma(c)+1$ / 计算项集的支持度 |
| 10: end for |
| 11: end for |
| 12: $F_k$ = $\{c \mid c \in C_k \wedge \sigma(c) \geq N \times minsup\}$ / 提取频繁 k-项集 |
| 14: result = $\cup F_k$ |

Apriori 算法的核心在于在枚举可能的事项集的同时对搜索空间进行剪枝，将不可能作为频繁项集的事项集去除。这部分的逻辑如算法 4-1 所示。算法中，Ck 表示候选 k-项集的集合，Fk 表示频繁 k-项集的集合。首先，初始化参数，若 Fk 不为空集，则扫描 1 次数据集，计算每个项的支持度。完成这一步后，便得到了所有频繁 1-项集的集合 F1（步骤 1 和步骤 2）。之后，使用之前得到的频繁(k-1)-项集来产生新的候选 k-项集（步骤 5）。随后，再次扫描数据集，通过计算候选项的支持度来确定样本集中包含在每个样本 t 中的 Ck 中的所有候选 k-项集（步骤 6~10）。最后，计算候选项集的支持

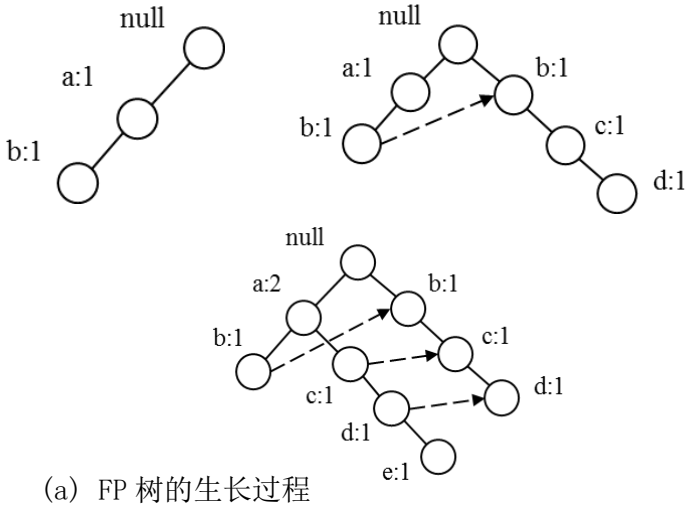
度，并选择支持度大于支持度阈值的所有候选集（步骤 12）。当迭代过程中没有新的事项集合产生时，Fk 为空，算法结束。

### 4.1.2　FP-growth

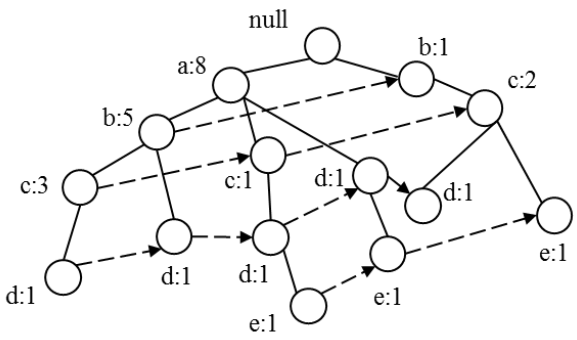
FP-growth 算法在 Apriori 算法之后提出。两种算法采用了截然不同的方法来优化关联规则发现的搜索空间[79]。如 4.1.1 节中所述，Apriori 算法采用“产生-测试”的范式，而 FP-growth 是利用 FP 树结构直接提取频繁项集。

FP 树是一种紧凑的数据结构，可以用于 4.1.1 节中所说的项集的格结构的压缩表示。具体地，把样本逐个读入 FP 树结构中，即将每个样本映射为树中的一条路径，如图 4.3 所示。在数据集中，不同样本可能会有一个或多个相同特征（事项），因此这些样本在 FP 树中的路径可能会有重叠部分。这种重叠部分就是 FP 树可以进行压缩的部分，因此，重叠部分越多，FP 树的压缩效果就越好。当 FP 树小到可以在内存中存储，就可以实现直接从内存中进行关联分析，这样减少了与硬盘交互的时间，使得挖掘频繁项集的效率倍增。为了说明 FP 树的结构，本文假设如下简单数据集，包含 10 个样本和 5 个项，图 4.3(a)展示了读入前 3 个样本后的 FP 树的生长过程。

| TID | 项 |
|---|---|
| 1 | {a, b} |
| 2 | {b, c, d} |
| 3 | {a, c, d, e} |
| 4 | {a, d, e} |
| 5 | {a, b, c} |
| 6 | {a, b, c, d} |
| 7 | {a} |
| 8 | {a, b, c} |
| 9 | {a, b, d} |
| 10 | {b, c, e} |




(a) FP 树的生长过程




(b) FP 树搜索过程

图 4.3　FP 树的结构

FP-growth 通过自底向上的顺序对项集进行探索，给定图 4.3(b)所示的树，首先查找以 e 结尾的频繁项集，然后依次查找 d, c, b， 最后是 a。每个样本都可以表示为 FP 树中的一条路径，所以，确定以 e 结尾的频繁事项集的问题可以转化为考察所有与节点 e 关联的路径。找到以 e 结尾的频繁项集后，再通过与节点 d 关联的路径来进一步寻找以 d 结尾的频繁项集，以此类推，直到考察了所有与节点 a，节点 b，节点 c 相关联的路径后，将各个节点对应的频繁项集进行汇总即可得到样本集中的频繁项集，也即强关联规则。

总结来说，FP-growth 是采用分治策略将一个问题分解为较小的子问题来进行强关联规则（频繁项集）的查找的。假设要发现包含事项 b 的强关联规则，需要先检查项集 b 是否为强关联规则。如果它是强关联规则，则考虑发现包含 ab 的强关联规则的子问题，接下来是 cb 和 db。依此类推就可以找到所有包含事项 b 的强关联规则。这种分治策略是 FP-growth 算法的关键策略。

FP-growth 在某些数据集的情况下，效率比标准的 Apriori 算法快几个数量级，不过它的运行效率依赖于数据集的压缩因子，如果生成的 FP 树叶子结点非常多，则算法性能会显著下降。否则，算法效率远高于 Apriori 算法。

## 4.2 地震事件集的定义

AETA 设备分布在北京、河北、四川、云南、西藏、广东、陕西、甘肃、台湾等多个地区，不同的地区地质结构、地震发生频率以及台站布设密度等条件均不同，这导致了不同地区地震发生前的前兆信号和信息也不同，这些变量给地震预测问题带来的困难已经超出本文的研究范围。为了控制变量，作者综合地震频率和 AETA 台站布设密度等因素，在全国范围内选取了四川部分地区和云南部分地区（经度为 97°E~ 109°E, 纬度为 25°N~ 35°N）作为研究区域。另外，AETA 系统从 2016 年 12 月逐步开始密集布设，直至 2017 年底初步完成密集布设，在这期间系统台站数量从 10 多台逐步增加到 200 余台，为了固定所研究台站的数量和地震时间集的数量，作者综合地震事件集的数量和 AETA 台站布设时间等因素，选择 2017 年 7 月 1 日~2019 年 3 月 1 日（共计 609 天）作为研究时间范围。综合地震事件集的数量和地震破坏性等因素，选择大于 4 级的地震作为地震事件。

根据这些标准，作者筛选出了 32 个符合条件的地震事件，在这个时间范围和区域范围内共有 38 个 AETA 台站。表 4.2 和 4.3 分别展示了部分地震事件和部分 AETA 台站信息。

表 4.2　地震事件统计表

| 时间 | 震级 | 经度 | 纬度 |
|---|---|---|---|
| 2017-07-02 | 4.1 | 101.71 | 25.12 |
| 2017-07-17 | 4.9 | 105.36 | 32.38 |
| 2017-08-08 | 7.0 | 103.82 | 33.20 |
| 2017-08-09 | 4.8 | 103.86 | 33.16 |
| 2017-08-10 | 4.3 | 103.85 | 33.16 |
| 2017-09-12 | 4.4 | 101.42 | 27.93 |
| 2017-09-30 | 5.4 | 105.00 | 32.27 |
| … | … | … | … |
| 2019-02-24 | 4.7 | 104.49 | 29.47 |
| 2019-02-25 | 4.9 | 104.50 | 29.48 |

表 4.3　AETA 台站信息

| 台站简称 | 经度 | 纬度 | 安装时间 |
|---|---|---|---|
| DJY | 103.65 | 30.98 | 2016-12-31 |
| KDGZ | 102.17 | 30.12 | 2016-12-24 |
| CX | 101.54 | 25.03 | 2016-12-30 |
| XC | 102.27 | 27.90 | 2016-12-19 |
| SMWJ | 102.28 | 29.31 | 2016-12-22 |
| SMSD | 102.35 | 29.23 | 2016-12-22 |
| … | … | … | … |
| JZG | 104.25 | 33.26 | 2017-06-13 |
| SP | 103.60 | 32.65 | 2017-06-16 |
| MC | 103.90 | 28.96 | 2017-06-14 |

根据上述表格数据，可以建立地震数据集，需要注意的是，本文在第三章中定义了地震事件时间窗口，为了增加正样本的数量，将这个时间窗口内的样本均视为正样本（地震数据集）。非地震数据集则为地震事件时间窗口外的样本。

## 4.3 频繁项集计算

频繁项集发现的前提是样本集中的特征都是离散化的特征，因为支持度的计算需要统计某一特征在正/负样本集中分别出现的次数，如果使用连续特征，出现相同数值的比例非常低，也就会是的特征的支持度非常低，难以发现频繁项集。本文使用式(4.1)对特征进行 n 段离散化：

$$q_i = \left\lfloor n \times \frac{p_i - \min(p_i)}{\max(p_i) - \min(p_i)} \right\rfloor \tag{4.1}$$

其中，⌊ ⌋为向下取整；$q_i$ 为离散化后的特征 q 的第 i 个值；$p_i$ 为连续特征 p 的第 i 个值；n 为离散化的分段数。表 4.4 和表 4.5 分别展示了进行离散化后的地震事件集和非地震事件集。

表 4.4 地震事件集

| | SRSS | right_high | … | middle_2_peak | SVD |
|---|---|---|---|---|---|
| 2017-07-01 | 3 | 4 | … | 1 | 5 |
| 2017-07-02 | 3 | 4 | … | 0 | 5 |
| 2017-08-01 | 5 | 3 | … | 2 | 5 |
| … | … | … | … | … | … |
| 2019-02-24 | 5 | 2 | | 0 | 5 |

表 4.5 非地震事件集

| | SRSS | right_high | … | middle_2_peak | SVD |
|---|---|---|---|---|---|
| 2017-07-03 | 3 | 4 | … | 1 | 3 |
| 2017-07-04 | 3 | 4 | … | 0 | 4 |
| 2017-07-05 | 5 | 3 | … | 2 | 2 |
| … | … | … | … | … | … |
| 2019-03-01 | 5 | 2 | | 0 | 1 |

本文使用 FP-growth 算法发现频繁项集，一般情况下，关联分析常被用于购物车分析，帮助商家找到消费者最喜欢组合购买的物品，方便商家进行促销活动，提高盈利。关联分析的主要目标是找到关联规则，并根据关联规则的支持度和置信度进行排序，从而找到出现最频繁的关联规则。对于一个关联规则 R:{A—>B}，支持度定义为事件 A 和 B 同时出现在总体事件集中的概率，而置信度定义为当 A 出现在事件集时 B 出现的概率。

以表 4.4 和表 4.5 中事件集为例，对于关联规则{SRSS=3—>right_high=4}，其在事件集 D 中的支持度定义为 SRSS=3 和 right_high=4 同时出现在事件集中的次数与 D 中事件总数之比，可以理解为在事件集 D 中关联规则{SRSS=3—>right_high=4}出现的概率。而关联规则{SRSS=3—>right_high=4}在事件集 D 中的置信度则定义为 SRSS=3 时 right_high=4 的概率。在本文研究的场景下，更加关心某个关联规则是否经常出现在地震事件集中或者是否经常出现在非地震事件集中，而不用太关心这个关联规则的项出现的条件概率，因此频繁项集的发现中，主要关注关联规则的支持度。本文中设定支持度阈值为 0.5，置信度阈值为 0.4，分别得到满足条件的地震事件频繁项集和非地震事件频繁项集。而在地震事件集的频繁项集中也可能存在非地震事件集的频繁项集，这说明这种情况下的频繁项集与地震事件基本没有相关性。因此，本文在得到的地震事件集的频繁项集中剔除掉非地震事件集的频繁项集。这样就得到了真正与地震事件相关的特征情况，从而可以从特征集中筛选出重要的特征，并且增强了模型的可解释性。

### 4.3.1　地震事件频繁项集

根据关联分析方法计算得到地震事件频繁项集，结果如表 4.6 所示。实验中取 n=5，m=7。

表 4.6　地震事件频繁项集

| 频繁项集 | 支持度($r_1$) |
|---|---|
| { (Mag_SRSS_num ,3)} | 0.74 |
| {(Mag_category_7,5)} | 0.67 |
| {( Mag_SVD_max,5),(Mag_STA_LTA,2)} | 0.62 |
| {(Sound_STA_LTA,0),(Eq_mag_max,5)} | 0.59 |
| {( Sound_category_1,1),(Eq_num,5),(Eq_mag_mean,5)} | 0.54 |
| {( Sound_category_0,0),(Mag_SVD_max,5),(Eq_mag_min,4)} | 0.52 |
| {( Sound_category_2,1),(Mag_category_7,5),(Eq_mag_mean,5)} | 0.47 |
| … | … |

### 4.3.2　非地震事件频繁项集

根据关联分析方法计算得到非地震时间频繁项集，结果如表 4.7 所示。实验中取 n=5，m=7。

表 4.7　非地震事件频繁项集

| 频繁项集 | 支持度($r_2$) |
|---|---|
| {((Mag_category_1,5))} | 0.86 |
| {(Mag_category_3,2)} | 0.77 |
| {( Mag_SRSS_acc_mean,0),(Eq_num,0)} | 0.72 |
| {(Eq_mag_mean,0),(Eq_mag_max,0)} | 0.67 |
| {( Sound_category_0,3),(Eq_num,0),( Eq_mag_mean,1)} | 0.67 |
| … | … |

## 4.4　特征降维

本文基于 AETA 系统的电磁扰动数据和地声数据提取了一元时间序列特征、多元时间序列特征、滑动时间窗口统计特征以及地震事件频率特征等 32 个特征。在这些特征中，有的与地震事件相关性高，有的则与地震事件相关性不高，这些特征进入模型后反而会降低模型的表现，导致模型过拟合。因此，通过一些有效的技术手段对特征空间进行选择和降维是一项有必要的工作。根据 4.3.1 节和 4.3.2 节中得到的地震事件频繁项集和非地震事件频繁项集可以发现，两个频繁项集中有部分重合的项集，例如，{( Sound_category_1,1),(Eq_num,5)}。这可能是由于地震频繁项集中包括了前兆信号和噪声信号，而非地震频繁项集包括了噪声信号和正常信号，重合的项集恰好由于噪声信号引起。

一种可行的特征选择/特征降维的方法是在地震事件的频繁项集中剔除掉所有在非地震事件的频繁项集，这样本文得到的频繁项集就是真正的地震前兆信号所具备的项集。得到做差的结果后，还可以进一步计算出不同台站对于地震事件的贡献度，从而可以在原始模型的基础上，对不同台站数据赋相应的权重，进一步优化模型。特征降维后的频繁项集结果见表 4.8。

表 4.8　特征降维后的频繁项集

| 频繁项集 | 支持度($r_1$-$r_2$) |
|---|---|
| { (Mag_SRSS_num,3)} | 0.24 |
| {( Sound_category_1,1),(Eq_num,5),( Eq_mag_mean,5)} | 0.21 |
| {(Eq_mag_max,5)} | 0.21 |
| {( Mag_SVD_max,5),(Mag_STA_LTA,2)} | 0.18 |

## 4.5　本章小结

本章提出了基于关联分析的特征选择方法。第一，分析了关联分析的概念和常用的 Apriori 算法和 FP-growth 算法。第二，提出了地震事件集的定义和非地震事件集的定义，并根据 AETA 系统的布设情况选取了研究区域范围和时间范围，在这个范围中得到了相应的地震事件集和非地震事件集。第三，在筛选得到的地震事件集和非地震事件集中利用 FP-growth 算法分别计算出两个事件集中支持度大于 0.5，置信度大于 0.4 的频繁项集。最后，找出在地震事件集的频繁项集而不在非地震事件集的频繁项集，得到与地震事件相关性最大的特征和频繁项集，为后续的模型提供选择之后的特征并提高模型的可解释性。

# 第五章　短临地震风险预测模型的建立与评估

上述章节所得出的样本集将作为本章建立模型的数据基础。根据本文所生成样本集的特点，选择决策树、支持向量机、梯度提升树算法，以研究范围内是否会有某级地震发生为目标建立短临地震风险模型。用混淆矩阵、AUC 等指标对模型进行评估。最后根据模型的预测结果对模型做出解释，以指导进一步的数据分析，地震预测工作。

## 5.1　方法对比与分析

### 5.1.1　决策树

决策树(Decision Tree, DT)是一种经典的解决分类与回归的方法[80]。决策树是一种自顶向下生长树形结构，恰好与第四章中所述的 FP 树相反，这样的生长过程就是对样本分类或者回归的过程。决策树模型的特点是模型简单，计算速度通常较快，而且决策树具有比较好的可解释性。但是，决策树对于数据的拟合程度通常没有复杂度较高的神经网络、集成学习模型等好。

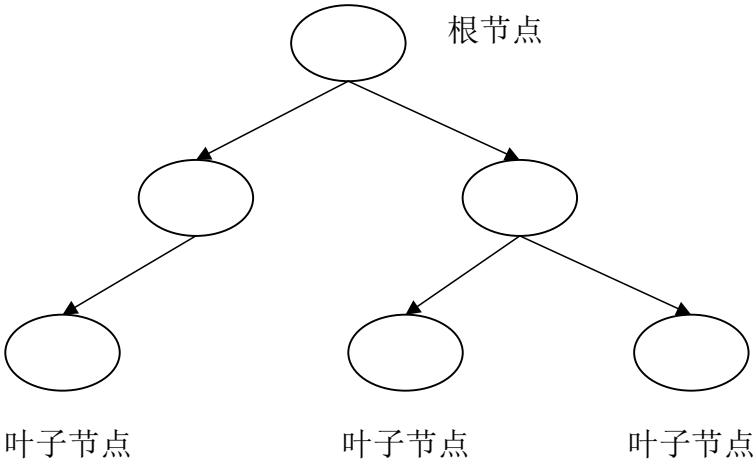



图 5.1　决策树

决策树学习的重要一步是分裂点选取，选取能够对当前数据集进行最佳划分的特征以及该特征的分裂点。通常分裂点的选取的准则是信息增益、信息增益比和基尼系数。

1. 信息增益和信息增益比

对于随机变量(X, Y)，其联合概率分布如式(5.1)所示：

$$P(X = x_i, Y = y_i) = p_{ij}, i = 1, 2, ..., n; j = 1, 2, ...m \tag{5.1}$$

条件熵 $H(Y|X)$ 表示在已知随机变量 $X$ 的条件下，随机变量 $Y$ 的不确定性，定义为 $X$ 给定条件下 $Y$ 的条件概率分布的熵对 $X$ 的数学期望，如式(5.2)所示：

$$H(Y|X)=\sum_{i=1}^{n}p_iH(Y|X=x_i) \tag{5.2}$$

其中，$p_i=P(X=x_i),i=1,2,...n$ 。

而信息增益表示在已知特征 $X$ 的前提下，类 $Y$ 的信息不确定度减少的程度。一般地，熵 $H(Y)$ 与条件熵 $H(Y|X)$ 之差即信息增益，如式(5.3)所示：

$$g(Y,X)=H(Y)-H(Y|X) \tag{5.3}$$

假设训练数据集标签向量为 $Y$，特征矩阵为 $X$，则 $X$ 的先验信息对数据集 $Y$ 的分类过程具有指示作用，这种指示作用会降低系统的不确定度。因此，一个特征的信息增益越大，说明该特征对当前分类任务的贡献越大，在决策树的 ID3 算法中，就采用信息增益作为查找分裂点的标准。

然而，需要注意的是，信息增益作为分裂点查找标准也是有其弊端的，一些学者发现一些取值非常多的特征很容易获得比较大的信息增益，从而使得决策树算法被诱导选取这些取值很多，但并没有为最后的标签预测做出匹配的贡献。例如，以 ID 作为一个特征，对于每个样本来说，ID 都是不同的，计算出来的 $H(Y|X)$ 会是 0，这样就会得到一个非常大的信息增益，而本文明确知道 ID 对于本文最后的分类而言是没有意义的。为了解决这个问题，就必须对取值过多的特征进行惩罚，学者们因此提出了信息增益比，如式(5.4)所示：

$$\mathrm{g}_R(\mathrm{Y},\mathrm{X})=\frac{g(Y,X)}{H_X(Y)} \tag{5.4}$$

其中，$H_X(Y)=-\sum_{i=1}^{n}\frac{|Y_i|}{|Y|}\log_2\frac{|Y_i|}{|Y|}$，$n$ 是特征 $X$ 的取值个数。这样，就可以很好地对取值过多的特征进行惩罚，从而真正选择到有意义的特征。

2. 基尼系数

一般来说，机器学习任务分为分类任务和回归任务 2 种。上述的 ID3 决策树和 C4.5 决策树主要是针对分类问题，而 CART 决策树既可以应用于分类问题，又可以应用于回归问题，主要原因是 CART 树是一个二叉树，通过递归的方法可以对连续特征进行划分，而 ID3 和 C4.5 均为多叉树，只能对离散特征进行划分。因此，CART 树在现代机器学习领域广受欢迎，诸如 GBDT、XgBoost、LightGBM 等梯度提升树都采用了 CART 树作为基学习器。而在 CART 分类树中，采用基尼系数来作为寻找分裂点的标准。假设样本集中有 $K$ 类，样本点属于第 $k$ 类的概率为 $p_k$，则基尼系数定义如式(5.5)所示：

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2 \tag{5.5}$$

对于给定的样本集 $D$，$c_k$ 是 $D$ 中属于第 $k$ 类的样本子集，则其基尼系数如式(5.6)所示：

$$Gini(D)=1-\sum_{k=1}^{K}(\frac{|c_k|}{|D|})^2 \tag{5.6}$$

当样本集 $D$ 根据特征 $A$ 是否取某值 $a$ 而被分为 $D_1$ 和 $D_2$ 两部分，那么在该条件下，集合 $D$ 的条件基尼系数如式(5.7)所示：

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \tag{5.7}$$

$Gini(D)$ 表示样本集 $D$ 的不确定性大小，$Gini(D,A)$ 表示在特征 $A=a$ 分割后的样本集 $D$ 的不确定性。

### 5.1.2 支持向量机

支持向量机(Support Vector Machines, SVM)是一类有监督机器学习算法，是由统计学习理论的创始人，前苏联科学家 V.Vapnik 于 20 世纪 60 年代提出[81]。开始，支持向量机只是作为线性分类器，后来经过 Bernhard E. Boser 等人的努力，通过核方法得到了非线性的 SVM，使得 SVM 可以处理在低维线性不可分的样本。与当下常用的机器学习算法，诸如梯度提升树、随机森林、神经网络等，不同的是，SVM 不需要大量样本来训练，因为它是根据样本中的少量支持向量来确定分类超平面。因此，当训练数据集较少的时候，SVM 也往往会带来不错的效果。

给定样本集 $D=(x_1,y_1),(x_2,y_2),...,(x_m,y_m)$，其中 $y_i\in\{-1,+1\}$，如上文所述，SVM 的目标就是在特征空间 $X$ 中找到一个超平面，使得正负样本到这个超平面的距离最大，这也就是 SVM 独有的间隔最大化。设分类超平面如式(5.8)所示：

$$W^Tx_i+b=0 \tag{5.8}$$

其中 W 为法向量，决定了分类超平面的方向，b 为平移量，决定超平面距离原点距离。对于样本 $(x_i,y_i)$，若使得式(5.9)成立，则样本集是线性可分的：

$$\begin{cases}W^Tx_i+b\geq+1 & y_i=+1\\ W^Tx_i+b\leq-1 & y_i=-1\end{cases} \tag{5.9}$$

此时的$W^T x_i + b = 0$就是一个分类超平面，使得等号成立的样本即为支持向量。但是，使得上式成立的超平面可能有很多，不是唯一解，SVM 的目标是找到使得正负样本间隔最大化的超平面。样本$(x_i, y_i)$到分类超平面的距离如式(5.10)所示：

$$r = \frac{W^T x_i + b}{\|W\|} \tag{5.10}$$

对于支持向量，有$W^T x_i + b = 1$，则最大化分类间隔的超平面需要满足式(5.11)和式(5.12)：

$$\max_{W,b} \frac{1}{\|W\|} \tag{5.11}$$

$$s.t.\ y_i(W^T x_i + b) \geq +1 \tag{5.12}$$

式(5.11)和式(5.12)可以通过拉格朗日乘子法对其对偶问题求解。具体细节本文不再展开叙述。

### 5.1.3　梯度提升树

梯度提升树（Gradient Boosting Decision Tree，GBDT）是一种应用广泛的集成学习方法[82]。提升树是以分类树或回归树为基本分类器的提升方法，被很多数据科学家认为是统计学习中性能最好的方法之一。目前所说的梯度提升树算法是基于 1996 年 Schapire 和 Freund 共同提出的 AdaBoost 算法演进而来的，二者的区别是 GBDT 限制了基学习器只能用 CART 决策树，但同属于 Boosting 算法。

Boosting 是一种采用加法模型的计算思想，是指通过串行的方法不断叠加弱分类器，通过学习弱分类器与真实标签之间的残差来不断增强类器的分类能力。假设前一轮迭代得到的分类器是$f_{t-1}(x)$损失函数为$L(y, f_{t-1}(x))$，本轮迭代的目标是找到一个弱分类器$h_t(x)$，使得本轮迭代的损失函数$L(y, f_t(x) = L(y, f_{t-1}(x) + h_t(x))$最小。举例来说，假设某人年龄为 30 岁，使用 GBDT 来拟合他的年龄，首先第一个学习器拟合得到 20 岁，发现与真正的年龄还差 10 岁，那么第二个学习器以 10 岁为目标继续拟合，得到 6 岁，此时距离真正的年龄还差 4 岁，第三轮的学习器以 4 岁为目标继续拟合，得到 3 岁，这时候三轮叠加起来的学习器拟合结果与真实年龄仅差 1 岁了，可以看到在迭代过程中，拟合误差不断减小，最后得到的强学习器的拟合能力远比每一轮的弱学习器强。

如上文所述，梯度提升树模型可以表示为决策树的加法模型：

$$f_M(x) = \sum_{m=1}^{M} T(x; \theta_m) \tag{5.13}$$

其中，$T(x;\theta_m)$表示决策树；$\theta_m$为决策树的参数；M 为树的个数。

梯度提升树算法采用前向分步算法，首先确定初始提升树 $f_0(x)=0$，则第 m 步的模型是：

$$f_m(x) = f_{m-1}(x) + T(x;\theta_m) \tag{5.14}$$

其中，$f_{m-1}(x)$ 为当前模型，通过经验风险极小化确定下一棵决策树的参数 $\theta_m$，

$$\theta_m = \arg\min_{\theta_m}(\sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i;\theta_m))) \tag{5.15}$$

通过式(5.15)可以确定梯度提升树的各个参数，得到预测模型。总结梯度提升树算法流程如下：

表 5.1　梯度提升树算法

**算法 5-1 梯度提升树算法**

**输入：**训练数据集 $T = \{(x_1, y_1),(x_2, y_2),...,(x_N, y_N)\}$

**输出：**提升树 $f_M(x)$

1. 初始化 $f_0(x) = 0$
2. 对 m=1,2,…,M
   a.计算残差 $r_{mi} = y_i - f_{m-1}(x_i)$，i=1,2,…N
   b.拟合残差 $r_{mi}$ 学习一个回归树，得到 $T(x;\theta_m)$
   c.更新 $f_m(x) = f_{m-1}(x_i) + T(x_i;\theta_m)$
3. 得到梯度提升树 $f_M(x) = \sum_{m=1}^{M} T(x;\theta_m)$

### 5.1.4　算法对比分析

上文中介绍的三种算法各有利弊，因此在需要谨慎研究算法在不同场景下的应用效果。首先，对比决策树和支持向量机这两种算法。这两种算法都是较早期的统计机器学习最常用的方法，受限于数据收集和处理技术，当时的统计机器学习方法都难以在大数据的场景下应用，常用于几百条到几万条的数据集中。在这个数量级的样本集中，决策树和支持向量机都不容易欠拟合，更不容易过拟合，两种方法的预测效果都比较理想。决策树的优势在于方法原理简单，计算复杂度低，在小批量数据上有着高速度和高精度的表现。而支持向量机的优势在于处理线性不可分样本的能力，它可以仅通过少量的支持向量样本对低维空间不可分的样本进行准确的划分。但是这种特殊的能力带来的是更复杂的算法和更高的计算复杂度。

接下来，对比梯度提升树与决策树、支持向量机算法。梯度提升树可以说是决策树的升级版，利用加法原理将多个决策树的预测结果叠加，从而获得更强的拟合、分类能力。在大数据样本集下，精度远高于决策树，但计算时间复杂度也高于决策树。在小型数据集和大数据集上，梯度提升树在精度和计算速度上均优于支持向量机，这主

要是由于支持向量机在原理上很难支持大数据集下的训练，虽然也经历了 J.Platt、T.Joachims、张学工等人的改进，但是支持向量机始终难以改善样本噪声给算法带来的影响。在小型数据集上，梯度提升树容易因为样本不足而出现过拟合问题，不过这个更多是数据集本身导致的问题而非算法的不足导致的问题，现有的梯度提升树的软件包都在工程和原理上进行了大量的优化，极力避免了过拟合问题。表 5.2 所示为三种算法的对比分析情况，其中 √表示具备该项能力或者该项表现优秀，O 表示该项表现一般，×表示不具备该项能力或者该项表现不好。

表 5.2　SVM、DT、GBDT 算法对比

| | 大型数据集 | | | 小型数据集 | | |
|---|---|---|---|---|---|---|
| | SVM | DT | GBDT | SVM | DT | GBDT |
| 学习能力 | × | O | √ | √ | √ | √ |
| 计算速度 | × | √ | √ | O | √ | √ |
| 回归问题 | √ | √ | √ | √ | √ | √ |
| 分类问题 | √ | √ | √ | √ | √ | √ |

除此之外，本文在 AETA 系统数据集上形成了小量样本集和全量样本集，分别对上述三种算法进行了实验验证。其中全量样本集共有 60900 条样本，小量样本集则在其中随机抽样 600 条样本，训练集和验证集样本比例为 5:1。样本空间的构建方式如第三章中所述，实验目标是预测研究区域内未来 7 天是否发生地震，为一个二分类问题。图 5.2 展示了在这个实验中决策树、支持向量机和梯度提升树三种算法在验证集上的 AUC 指标。

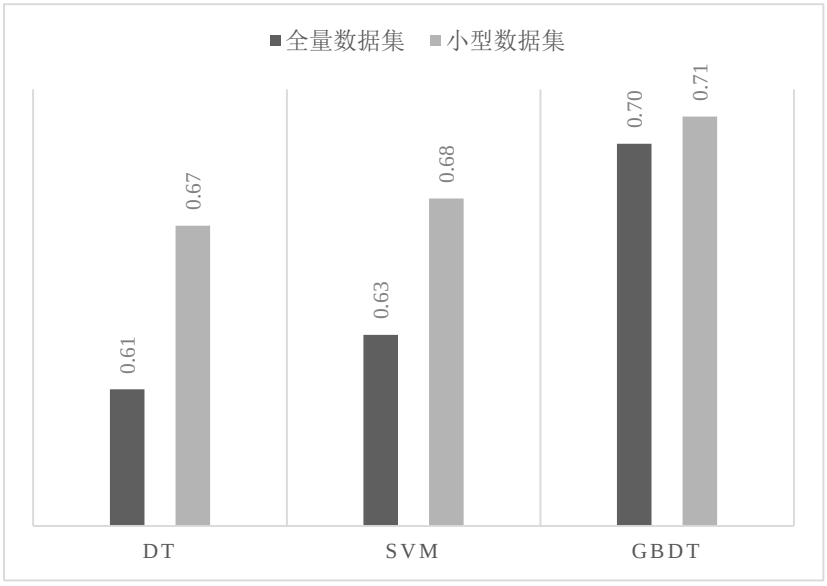



图 5.2　算法 AUC 对比结果

可以看到，实验结果与上述理论分析结果相吻合。GBDT 在全量数据集合小数据集上的表现均优于其他两个算法。因此在本文的研究中选则 GBDT 作为短临地震风险预测模型的算法。

## 5.2 模型的评价指标

在统计机器学习中，模型评价指标是直接决定一个算法或者预测效果好坏的标准。如前文中所述，统计机器学习一般可分为回归和分类两种任务，这两种不同的任务对应有不同的评价指标[83]。

对于回归任务，常见的多是基于距离的评价指标，例如均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)等，具体计算方式不同，但原理都是用于衡量模型预测结果 $\hat{y}$ 与真实结果 $y$ 之间的距离误差。

对于分类问题，常见的评价指标有准确率(accuracy)、查准率(precision)、查全率(recall)、$F_1$ 得分($F_1$-score)、AUC 等等。具体的，对于二分类问题，存在 4 种情况，将这 4 种情况出现的总数分别记作：

TP：将正样本预测为正样本数；

FN：将正样本预测为负样本数；

FP：将负样本预测为正样本数；

TN：将负样本预测为负样本数。

那么，查准率定义为：

$$P = \frac{TP}{TP + FP} \tag{5.16}$$

查全率定义为：

$$R = \frac{TP}{TP + FN} \tag{5.17}$$

$F_1$ 得分定义为查准率和查全率的调和均值：

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{5.18}$$

现实情况中，仅通过查全率、查准率往往不足以全面评价一个分类器的表现，原因是分类问题一般得到的预测值都是一个在[0,1]区间的值，表示分类器认为某样本属于某类的概率，但是并不表示这个值大于 0.5 就说明该样本属于这一类，而是需要一个人为设定的阈值来判断样本最终的类别归属，而这个阈值往往直接影响查全率、查准率等指标。这种情况下，相比查准率、查全率等指标，AUC 更能反映出分类器的真实表现。AUC 是受试者工作特征(ROC)曲线下的面积，也表示正样本排在负样本之前的概率，是一个在[0,1]区间的值，AUC 的值越大，说明分类器越能将正负样本区分开来。

ROC 曲线是对学习器给出的样本预测的结果进行排序，按这个顺序逐个把样本作为正例进行预测，每次计算出真正率(True Positive Rate, TPR)和假正率(False Positive Rate, FPR)的值，并分别作为纵轴和横轴于坐标系中画出，最后按顺序将每个点连接起来。图 5.3 展示了 ROC 曲线，该曲线下面积 AUC 的值为 0.71。

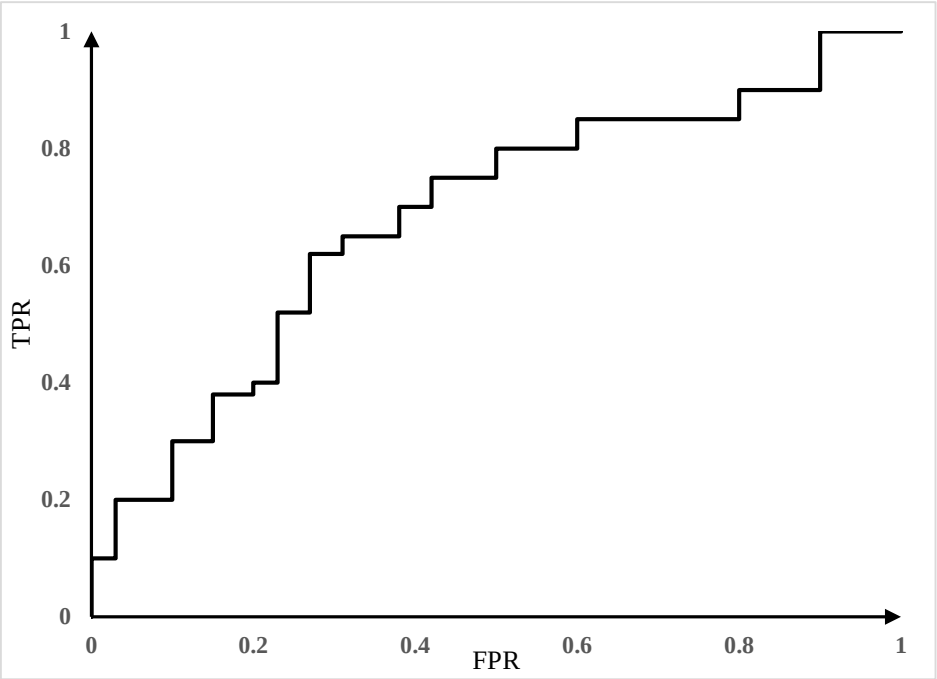



图 5.3　ROC 曲线示意图

在多分类问题中，评价指标依然可以通过将样本分为本类-其他类的形式而继续沿用二分类问题评价指标，但是当类别较多时，评价指标也随之增多，研究者并不能很容易地通过多个指标来合理评估模型表现。因此，实际情况中往往通过混淆矩阵(Confusion Matrix, CM)来评价多分类问题的模型表现。

由于本文所研究的是一个三目标（发震时间、发震地点、发震震级）预测问题，所以在具体的模型评价上，不能直接照搬上节中所述的统计机器学习中的评价标准，需要切合目标来定义模型评价标准。

根据作者对地震事件预测的理解，认为虽然时间、地点、震级在地震事件的描述和预测中都至关重要，但是在本文的研究框架下，这三要素中，时间的重要性要高于其他两个要素。因此，在评价体系中，时间要素占据最大的权重。又由于本文的研究框架事先选择了一个比较大的区域（如第三章中“震中范围选取”中所述）作为研究区域，震中区域已经有所限制，故而，震级的重要性大于地点的重要性。

基于此，按第三章中“时间窗口选取”所述，作者认为，当模型预测未来的时间窗口内会发生地震，且未来时间窗口内确实有地震，则算作一次发震时间准报，否则发震算作误报。在发震时间准报的前提下，震级预测正确，则算作一次发震时间和震

级准报，否则算作误报。在发震时间和震级准报的前提下，地震区域预测正确，则算作一次三要素准报，否则算作误报。

## 5.3 地震事件时间和震级预测子模型

AETA 系统在区域内布设密度不均匀以及地震事件数量不足使得地震发生地点难以预测。因此，本文综合台站布设密度和地震事件数量来选择研究区域，尽可能在布设密度比较均匀且地震事件发生较频繁的区域进行模型研究。根据统计，在系统布设范围内，东经 97°E~ 109°E, 北纬 25°N~ 35°N 这个区域满足上述条件。所以本文先选择这个区域研究该区域内的地震时间和地震震级预测模型。因为在地理上已经确定了一个大区域，所以可以通过控制变量法，暂时忽略对于地震震中的预测。

### 5.3.1 时间窗口对模型的影响

如第三章中“时间窗口的选取”中所讨论的，现有的研究中仅指出了短临地震预测的时间窗口为数天至数十天。在本文的研究框架中，希望能够得到一个确定的时间窗口，但是时间窗口的大小变化也同时影响着样本的标注结果。换言之，本文的研究框架中，样本空间是时间窗口大小的函数。这样，对于时间窗口 $n \in [1,15]$，本文分别生成了样本集 $D_i, i = 1,2,...,15$。为了突出时间窗口对模型的影响，暂时将问题退化为一个二分类问题，即给定样本 $x_j \in D_i, i = 1,2,...,15, j = 1,2,...,N$ (N 为样本数量)，预测未来 n 天内是否有地震发生。据此，可以根据不用样本集训练 15 个不同的模型，通过对比这 15 个模型在验证集上的 AUC 指标来最终确定时间窗口。

实验参数如下表所示：

表 5.3 实验参数表 1

| 参数名称 | 参数值 |
|---|---|
| 模型: | GBDT |
| 研究区域: | [97,109] [25,35] |
| 研究时间: | 20170701-20190301 |
| 样本大小: | 609 |
| 训练集大小: | 502 |
| 验证集大小: | 107 |

实验结果如图 5.4 所示，可以看到当 n=7 时，模型的 AUC 值最高（0.75），而当 n=1 时，模型的 AUC 值最低。所以，作者认为，在该研究区域（东经 97°E~ 109°E, 北纬 25°N~ 35°N）内，取滑动窗口为 7 天时，地震事件时间预测效果最优。这样，

就将问题转化为了预测未来 7 天，研究区域(东经 97°E~ 109°E, 北纬 25°N~ 35°N)内是否会发生地震。

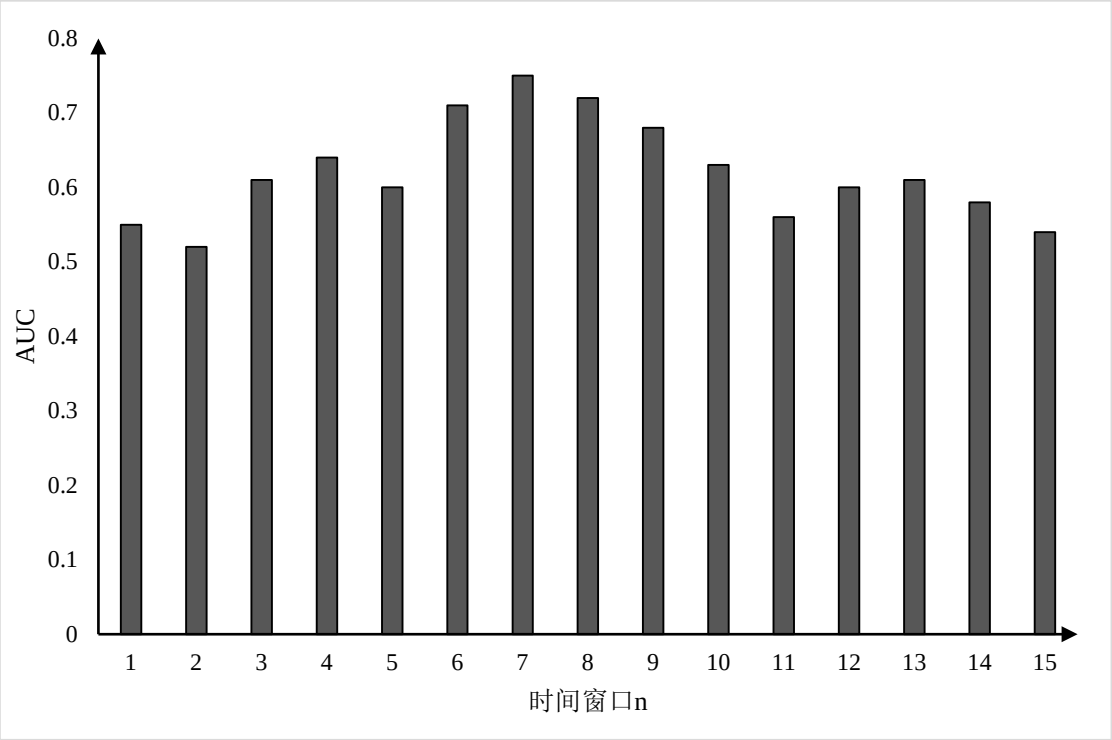



图 5.4　时间窗口 n 与 AUC 之间的关系

### 5.3.2　风险指数与发震阈值的确定

在 5.3.1 小节中，确定了预测模型的时间窗口为 7 天，对于震级预测只是按照有地震和无地震粗略地划分。本小节中，主要讨论地震震级的预测。

据本文统计，自 2017 年 7 月 AETA 系统设备大规模布设以来，研究区域（东经 97°E~ 109°E, 北纬 25°N~ 35°N）内的有感地震(>3.0 级)一共发生了 186 次，震级范围在[3.0,7.0]之间。中位数为 3.3 级，四分之三分为点为 3.7 级，这说明在这 186 次地震中 75%都小于 3.7 级，而在区间[3.7,7.0]中，只有 46 个地震，这个数量对于高精度的震级预测来说显然是不足够的。

为此，本文将地震事件按震级分桶，将本属于回归问题的地震震级预测问题退化为一个分类问题。数据分桶情况如下：[0,3]区间为无地震；[3,4]区间为无风险地震；[4,7]区间为有风险地震。在 5.3.1 小节中所述的二分类模型的基础上，加上分桶的震级样本，训练一个多分类器，便得到了在研究区域（东经 97°E~ 109°E, 北纬 25°N~ 35°N）内的地震事件时间、震级预测模型。取地震事件时间窗口 n=7，在与 5.3.1 小节中的实验参数相同的情况下，得到模型的混淆矩阵如图 5.5 所示：

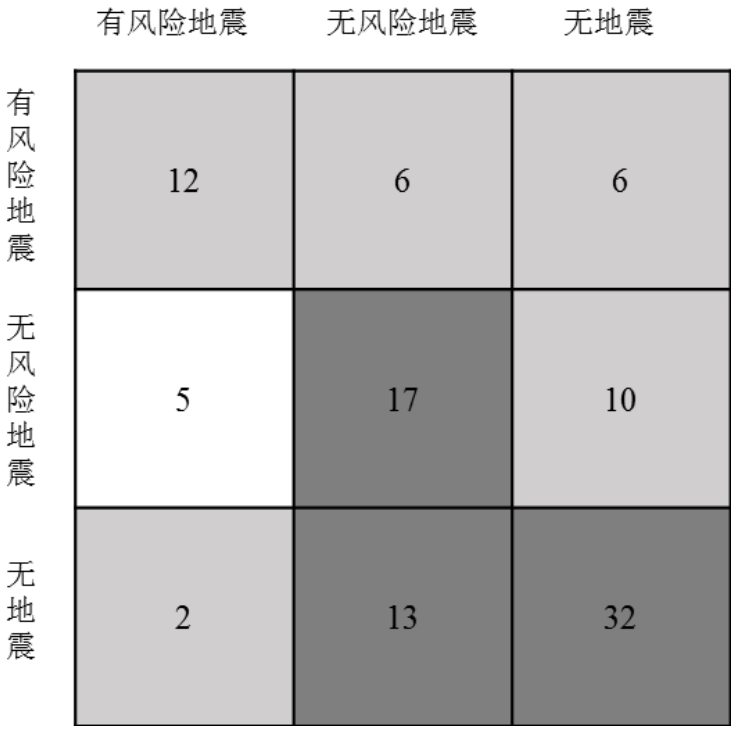



图 5.5　地震时间、震级预测混淆矩阵

模型对于无地震、无风险地震、有风险地震的预测查准率分别达到 0.50,0.53,0.68；查全率分别达到 0.67, 0.48, 0.63。可以看到，虽然在 5.3.1 小节中的 AUC 值可以达到 0.75，看上去是一个不错的模型，但是加上震级预测后的混淆矩阵中，有风险地震的预测准确率较低，而无地震和无风险地震的预测准确率较高。原因可能是由于无地震事件的样本明显多于地震事件样本，从而导致在模型训练过程中，为了降低损失函数无形中提升了无地震事件样本的权重。此外，由于模型同时预测了地震震级和地震时间两个目标，5.2 节中所述的模型评价标准也会相应变得更加苛刻，所以导致了预测地震震级和时间上模型的准确率降低。

## 5.4　基于 AETA 的短临地震风险预测模型

上一节中介绍的地震时间、震级预测模型的适用范围是在东经 97°E~ 109°E, 北纬 25°N~ 35°N 这个区域内，对于地震事件地点的预测而言，这个范围过于宽泛了。受到 Phoebe M. R. DeVries 等人余震预测工作的启发，本文将研究区域以 1 度经度和 1 度纬度进行进一步细分，在此区域内，每个网格的真实大小约为 111km*97km。

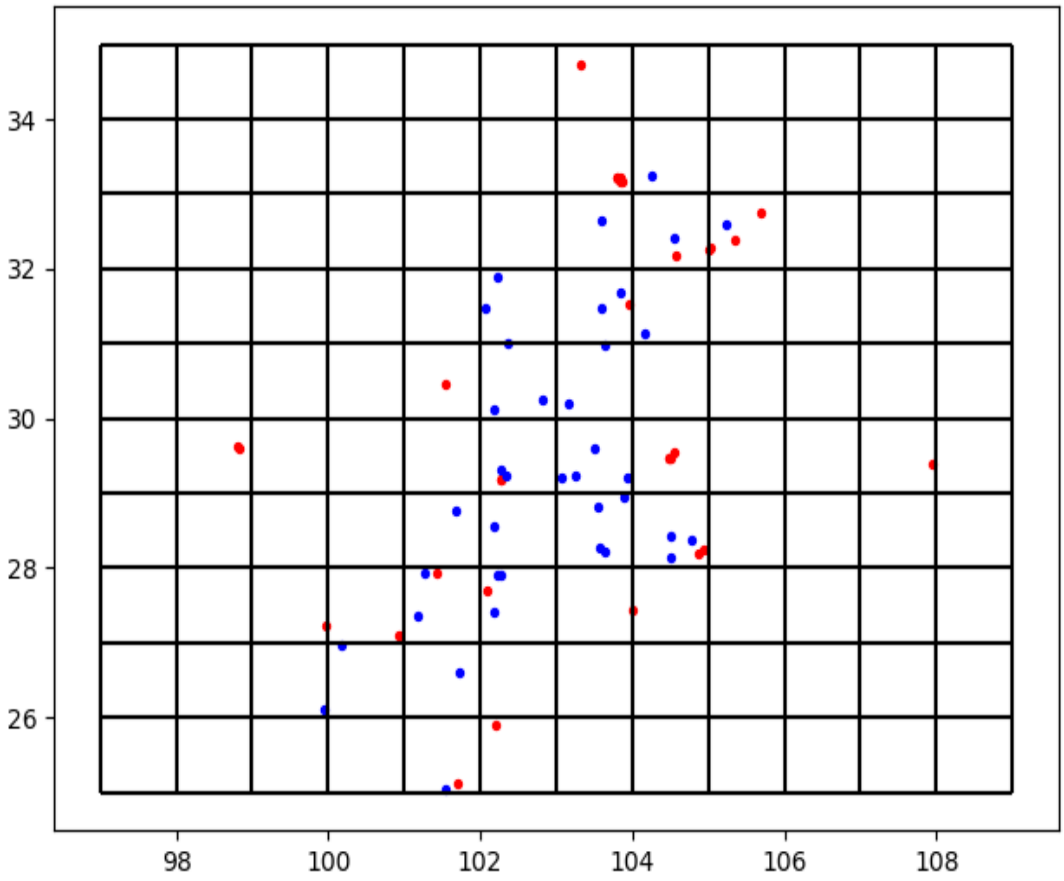



图 5.6　区域划分结果

如图 5.6 所示，本文将研究区域进行细分后发现在一些小区域中没有 AETA 台站，或者距离 AETA 台站较远，相较于区域中有 AETA 台站或者距离 AETA 台站较近的小区域，无疑缺失了信息。为了减少这种信息缺失给研究带来的影响，本文制定如下策略：

1. 遍历所有小区域，将存在 AETA 台站的小区域标记为 $A_1$；不存在 AETA 台站但区域中心距离最近的 AETA 台站小于等于 200 千米的区域标记为 $A_2$；不存在 AETA 台站且区域中心距离最近的 AETA 台站大于 200 千米。
2. 剔除属于 $A_3$ 的区域，在剩下的两种区域中按第三章所述特征空间建立样本。将原本一个大区域的样本转换为每个小区域的样本并加入每个小区域的地理信息。
3. 根据上述策略重新生成细分区域样本。

得到细分区域样本后，进行数据实验，相关实验参数如下表所示：

表 5.4　实验参数表 2

| 参数名称 | 参数值 |
|---|---|
| 模型: | GBDT |
| 研究区域: | [97,109]*[25,35] |
| 研究时间: | 20170701-20190301 |
| 样本大小: | 50200 |
| 训练集大小: | 45040 |
| 验证集大小: | 5160 |

本文在细分区域后的验证集上得到 AUC 指标为 0.72，混淆矩阵如表 5.5 所示：

表 5.5　短临地震风险预测模型混淆矩阵

| | | 真实标签 | | |
|---|---|---|---|---|
| | | 无地震 | 3-4 级地震 | 4-7 级地震 |
| 模型预测 | 无地震 | 2877 | 454 | 187 |
| | 3-4 级地震 | 434 | 853 | 65 |
| | 4-7 级地震 | 98 | 134 | 58 |

从表 5.5 中可以看到，精细划分区域后的地震三要素模型对于无地震事件、3-4 级地震、4-7 级地震的查全率分别为 0.84，0.59，0.19；对于无地震事件、3-4 级地震、4-7 级地震的查准率分别为 0.81，0.63，0.20。可以看到，加入细分区域后对于无地震事件的查准率和查全率相比时间和震级预测子模型的查准率和查全率有一定提高，而 3-4 级地震和 4-7 级地震的预测查准率和查全率有所降低。其原因是，细分区域后，样本集得到了扩充，而扩充的多数样本是无地震样本，少数样本是 3-4 级地震，极少数样本是 4-7 级地震。在模型训练过程中，对于无地震事件的学习更加充分，因此，对无地震事件的预测结果有所提升。增加的少数有地震样本受到大量无地震事件的干扰，在训练中的效果则有所下降。为了更直观地理解这个结果，本文将所做的研究与其他地震预测相关的工作进行对比。如表 5.6 所示。

表 5.6　本文工作与其他相关工作对比表

| | 研究信号 | 时间窗口 | 震中误差 | 震例 | 结果 |
|---|---|---|---|---|---|
| 地磁日变化畸变[84] | 地磁信号 | 90 天 | 1000km | 8 | 准确率 30%-60% |
| 危险理论[85] | 历史地震 | 5~15 天 | — | — | 准确率 52%，误报率 30% |
| 地磁低点位移法[86] | 地磁信号 | 90 天 | 300km | 246 | 查全率 16%，查准率 38% |
| 加卸载响应比[27] | 地下应力 | — | 300km | 9 | 查全率 78%，查准率— |

续表 5.6　本文工作与其他相关工作对比表

| | | | | | |
|---|---|---|---|---|---|
| 本文工作 | 电磁、地声、历史地震 | 7 天 | 111km | 186 | 无地震，3-4 级地震，4-7 级地震查全率和查准率分别为 84%，59%，19%；81%，63%，20%。 |

## 5.5　模型的优化及改进的讨论

本文从 AETA 系统数据出发，根据统计机器学习理论，使用梯度提升树所建立的短临地震风险预测模型在地震时间、地震震级、地震震中的预测结果中 AUC 指标都超过了 0.7，有一定有效性。但是具体到混淆矩阵中的指标表现明显低于 AUC，因此，本文提出的建模体系仍然有提升空间。本小节中主要从模型的可扩展性和主震和余震对模型的影响角度，讨论模型的优化和改进方向。

### 5.5.1　模型可扩展性

AETA 系统设备仍然在不断增加，新增的台站可以为地震预测提供更多的地震前兆信息，从而可以进一步提升模型预测的表现。在本文的研究框架中，样本空间是根据细分的小区域来定义的，在一个小区域内的样本仅由与其 200 公里内的台站信息组合而成。因此，在这个框架下，新增加的台站在积累一定时间的数据后，可以加入所属小区域的样本中。不过，在加入时需要重新对该区域的样本聚合方式进行修正。通过这样的策略，可以使得 AETA 系统新安装设备不断加入现有的预测模型中，随着时间和数据的积累，模型的预测效果也可以得到一定的提升。

从模型的泛化能力来说，本研究中的样本全部来源于东经 97°E~109°E，北纬 25°N~35°N 区域内的 38 个 AETA 台站以及该区域内的 186 个历史地震事件。在所研究区域和时间范围内，对验证集中的 5160 条地震/非地震样本进行验证，得到了如表 5.4 所示的混淆矩阵，因此，本模型在该区域内的数据集上具有一定的泛化能力。由于地震与震源地下结构复杂的相关性仍然没有明确的理论研究，AETA 台站在不同区域布设的密度也不同，本模型所学习到的参数难以直接应用于其他区域，但是本研究提出的这种预测研究框架完全可以在其他区域数据集上进行迁移学习，对新区域数据集的再训练，可以使得模型具有更好的可扩展性。

### 5.5.2　主震和余震对模型的影响

一般来说，在一个大地震往往会在附近区域出现一系列中强地震，而大地震和这一系列的余震之间的时间相关性、震中相关性往往较大，而且后续余震的前兆信号也

有可能被主震的前兆信号所覆盖。因此，模型对于余震的预测准确率会低于主震，从而在整体上降低了模型表现。在 5.4 节中建立的短临地震风险预测模型中，没有区分主震和余震，若能将主震和余震分开讨论，有可能进一步提升预测模型的表现。本文设计了如下策略来判别主震和余震。

1．在研究区域和研究时间范围内，找出所有地震事件。

2．在地震事件列表中，找到大于 5 级的地震事件 A，若该地震发生后 10 天内，相同地点又发生了小于 5 级的地震事件 B，则认为地震 B 是地震 A 的余震。

通过该策略，可以将地震事件集分为主震事件集和余震事件集，在后续的研究中，可以继续探讨，以期望进一步提升该模型的预测效果。

### 5.5.3 地震发生机理的探讨

至今为止，对于地震的发生机理尚存争议，虽然学术界提出了页岩动态应力、断层滑动等假说，但是依然没有形成一个同时具备理论依据和实验数据支持的结论。本研究中心结合 AETA 系统布设范围内的地震事件和观测数据认为，地震在发生时是一个能量聚集过程，在这个过程中，有可能会冲破地壳从而引起地震。在文中所提到的电磁数据 SRSS 波的出现和消失便是体现能量聚集过程的一个现象。然而，目前由于观测数据尚不足够，仍需要在布设范围内积累更多的地震事件，尤其是大地震，来证明该假设。在当前机理尚不明确的情况下，通过观测数据来间接的预测地震事件的三要素是一种可行的方法，但并不是最直观最准确的方法。随着数据的积累和理论的发展，对于地震事件的预测准确率会进一步提高。

## 5.6 本章小结

本节基于 AETA 系统数据，利用统计机器学习方法，建立了短临地震风险预测模型对地震的发生时间、震级、震中三要素进行预测，并对模型进行评估。首先，介绍了统计机器学习中常用的算法，对决策树、支持向量机、梯度提升树在 AETA 系统数据中的表现做出对比，并最终选择梯度提升树作为模型。其次，分别对地震事件的时间、地点和震级三要素的预测展开进行讨论，得到了合理的实验参数。然后，在该实验参数下，对东经 97°E~ 109°E, 北纬 25°N~ 35°N 区域内，2017 年 7 月 1 日至 2019 年 3 月 1 日之间的地震事件进行数据实验。在验证集中模型对于无地震事件、3-4 级地震、4-7 级地震的查准率分别为 0.81，0.63，0.20；查全率分别为 0.84，0.59，0.19。最后，从模型的可扩展性和主震余震对模型的影响两个方面讨论了模型的优化和改进方向。

# 第六章　总结与展望

## 6.1　总结

本文基于多分量地震监测系统 AETA 的数据，利用统计学习方法、机器学习算法建立模型，对地震的时间、地点、震级三要素进行预测。为了达成地震三要素预测的目标，本文主要从以下 4 个方面开展研究：

1．总结地震预测问题的研究现状。对现有的地震监测预测方式方法进行介绍，并总结其优劣性，重点介绍了基于统计学习方法和机器学习算法的地震预测研究。

2．AETA 系统电磁扰动原始数据、地声原始数据的分析以及预处理。AETA 系统数据积累量达 18TB，数据中蕴含着丰富的地震信息。本文对原始数据在时域和频域上分析，从时域上，利用 STA-LTA 算法对数据的幅值、均值、方差等特征进行分析；频域上，利用快速傅里叶变换对数据的频谱特性进行分析。在数据预处理方面，提出了 AETA 系统缺失数据的补全方法。

3．研究了基于 AETA 系统数据的特征空间生成方法。使用时间序列数据挖掘相关方法，对 AETA 系统单个台站的时间序列数据进行模式识别、序列描述；使用奇异值分解对 AETA 系统某个区域内多个台站的时间序列数据进行分析。基于统计方法，对地震的时间范围、地点范围、震级范围进行研究，并得出样本标注方法。

4．研究了建立基于决策树模型的地震三要素预测模型，并设计相应数据进行对比实验，通过 AUC、准确率、查全率等统计指标对模型的有效性进行检验。

在 AETA 系统数据的特征空间生成方法上，本文提出了以天为周期，对区域划分生成样本空间的方法，该方法避免了不同台站对同一地震事件重复采样而导致冗余特征的产生，同时又通过对区域细分增加了有效样本，弥补了样本量较少的劣势。本文利用该方法提取出了 Mag_SRSS_num、Mag_SVD_max 等 15 个有效特征。

在短临地震风险预测模型的建立与评估中，本文首先通过交叉验证得到当地震时间窗口 n=7 时，模型对于地震预测的效果最优，从而确定了地震预测的时间窗口。其次，通过将震级离散化，区域细分等方法，将震级和区域的预测转化为分类问题，从而可以在经纬度 1°的范围内对震中进行预测。最后，选取东经 97°E~ 109°E, 北纬 25°N~ 35°N 区域内的 186 个地震事件(>Ms3.0)以及 38 个 AETA 台站在 2017 年 7 月 1 日至 2019 年 3 月 1 日期间的数据作为实验对象进行数据实验。地震时间震级预测子模型对无地震、3~4 级地震、4~7 级地震的查准率分别达到 0.50,0.53,0.68；查全率分别达到 0.67, 0.48, 0.63，短临地震风险预测模型对于无地震事件、3-4 级地震、4-7 级地震的查准率分别为 0.81，0.63，0.20；查全率分别为 0.84，0.59，0.19。

## 6.2 展望

本文提出的模型在地震风险的预测中具有一定效果，但是由于地震事件本身的形成机理尚不清晰以及 AETA 系统的数据量仍需进一步积累。因此，现阶段的研究成果仍难难以直接应用于实际预测，本文所做的工作仅仅为地震预测这一科学界难题提供了一种可能的解决思路和研究框架，希望后来的研究者能够从中获取一丝灵感。基于本文所做的微小的工作，作者认为在未来，可行的提升地震预测的思路和方向有如下：

1. AETA 系统完成大密度布设至今还不足两年，这期间，可供研究的大地震仅有 1 例。在可以看到的未来，随着 AETA 系统日臻完善，积累的大地震数据逐渐增加，会使得地震预测效果有所提升。

2.主震和余震分开进行研究以及模型的可扩展性研究，可能有助于提升现有模型的预测结果，可以就这两个方向进行更深入的研究和讨论。

# 参考文献


[1] 陈运泰.地震预测:回顾与展望[J].中国科学(D 辑:地球科学),2009,39(12):1633-1658.

[2] 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, 28(2):1-24.

[3] Zhu Z , Toksoz. Seismoelectric and seismomagnetic measurements in fractured borehole models[J]. GEOPHYSICS, 2005, 70(4):F45-F51.

[4] 蒋长胜，吴忠良．关于中强震前的应变加速释放现象[J]. 中国科学院大学学报，2005, 22(3):286-291.

[5] 朱治国，王晓强，刘代芹等．2005～2009 年喀什—伽师地区重力场变化与地震[J]. 地震研究, 2011, 34(2):143-147.

[6] Zhu Y, An X. Variation of gravity feild before and after PanZhihua Ms6.1 and YaoAn Ms6.0 earthquakes[J]. Journal of Geodesy & Geodynamics, 2010,30(04):8-11.

[7] 几次大震前的地面和空间电磁场变化[J]. 地球物理学报, 2011, 54(11):2885-2897.

[8] Gershenzon N I, Gokhberg M B, Yunga S L. On the electromagnetic field of an earthquake focus[J]. Physics of the Earth & Planetary Interiors, 1993, 77(77):13-19.

[9] Zhao G Z, Yaxin B I, Wang L F, et al. Advances in alternating electromagnetic field data processing for earthquake monitoring in China[J]. Science China Earth Sciences, 2015, 58(2):172-182.

[10] 朱日祥，刘青松，郭斌．近 12000 年以来北京地区地球磁场变化机理探讨[J]. 地球物理学报, 2001, 44(2):211-218.

[11] 倪喆．洱源 5.5 级地震前后地磁场变化异常特征分析[J]. 地震研究, 2014, 37(3):426-432.

[12] Utada H, Shimizu H, Ogawa T, et al. Geomagnetic field changes in response to the 2011 off the Pacific Coast of Tohoku Earthquake and Tsunami[J]. Earth & Planetary Science Letters, 2011, 311(1-2):11-27.

[13] Mao T E, Wang T C, Yao J L, et al. The variations of the degree of ground resistivity anisotropy during the tangshan earthquake[J]. Acta Seismologica Sinica, 1995, 8(4):621-627.

[14] 钱复业，赵丰林．地震前地电阻率的异常变化[J]. 中国科学化学:中国科学, 1982, 12(9):831-839.

[15] 杜学彬，薛顺章，郝臻等．地电阻率中短期异常与地震的关系[J]. 地震学报，2000, 22(4):368-376.

[16] 常祖峰，谢阳，常昊．2018 年景洪 M4.9 地震地下水前兆异常特征[J]. 国际地震动态，2018, No.476(08):119-120.

[17] Satake H, Murata M, Hayashi H. Chemical characteristics of groundwater around the Mozumi-Sukenobu fault and the implication for fault activity[J]. Geophysical Research Letters, 2003, 30(7):8-10.

[18] 崔月菊，杜建国，陈志等．2010 年玉树 Ms7.1 地震前后大气物理化学遥感信息[J]. 地球科学进展, 2011, 26(7):787-794.

[19] Schnell R C, Cunningham M C, Vasel B A, et al. Atmospheric Baseline Monitoring Data Losses Due to the Samoa Earthquake[C]. //American Geophysical Union,2009.

[20] Hasegawa A , Yoshida K , Asano Y , et al. Change in stress field after the 2011 great Tohoku-Oki earthquake[J]. Earth and Planetary Science Letters, 2012, s 355–356:231–243.

[21] Skordas E S, Sarlis N V. On the anomalous changes of seismicity and geomagnetic field prior to the 2011 MwMw 9.0 Tohoku earthquake[J]. Journal of Asian Earth Sciences, 2014, 80(2):161-164.

[22] 张学民, 钱家栋, 欧阳新艳等. 新疆于田 7.2 级地震前的电离层电磁扰动[J]. 空间科学学报, 2009, 29(2):213-221.

[23] 马钦忠, 方国庆, 李伟等. 芦山 M_S7.0 地震前的电磁异常信号[J]. 地震学报, 2013, 35(5):717-730.

[24] Zhang X M, Liu J, Qian J D, et al. Ionospheric electromagnetic disturbance before Gaize earthquake with M_S6.9,Tibet[J]. Earthquake, 2008(03):14-22.

[25] 张克亮, 马瑾, 魏东平. 超导重力仪检测 2011 年日本东北 M_w9.0 级地震前的重力扰动信号[J]. 地球物理学报, 2013, 56(7):2292-2302.

[26] 李璐. 台阵处理技术和模板匹配滤波技术在微弱地震信号检测中的应用[J]. 国际地震动态, 2017(03):38-39.

[27] 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]. 中国科学化学: 中国科学, 1991, 21(5):512-518.

[28] 尹祥础, 陈学忠, 宋治平等. 响应比理论用于地震预报的进展[J]. 地震, 1994(s1):18-24.

[29] Zhang W J, Chen Y M, Zhan L T. Loading/Unloading response ratio theory applied in predicting deep-seated landslides triggering[J]. Engineering Geology, 2006, 82(4):234-240.

[30] 龙锋, 蒋长胜, 冯建刚等. 历史大地震破裂区地震危险性的地震活动性定量分析——以南北地震带中北段为例[J]. 地震, 2012, 32(3):98-108.

[31] 邵延秀, 袁道阳, 梁明剑. 滇西南地区龙陵-澜沧断裂带地震危险性评价[J]. 地震学报, 2015(6):1011-1023.

[32] 刘文龙, 王金周, 陈宇卫等. 研究地震空区及条带附近地震破裂特征方法的误差讨论[J]. 地震工程学报, 2001, 23(3):224-229.

[33] Toda S, Stein R S, Beroza G C, et al. Aftershocks halted by static stress shadows[J]. Nature Geoscience, 2012, 5(6):410-413.

[34] Hirose S, Toda S. Stress shadow effect found in recent seismic sequences associated with Japan's large earthquakes[J]. Agu Fall Meeting Abstracts, 2010.

[35] 林邦慧, 胡小幸. 1975 年 2 月 4 日海城地震的破裂过程与海城地震序列空间分布的图象[J]. 中国地震, 1988(2):51-60.

[36] Cimellaro G P, Marasco S. Earthquake Prediction[J]. Earth-Science Reviews, 2018, 12(3624):1364.

[37] 张国民, 钮凤林, 邵志刚. 帕克菲尔德地震预报实验场:2004 年 6 级地震及其对地震物理和地震预测研究的影响[J]. 中国地震, 2009, 25(4):345-355.

[38] Johnston M J S, Sasai Y, Egbert G D, et al. Seismomagnetic Effects from the Long-Awaited 28 September 2004 M 6.0 Parkfield Earthquake[J]. Bulletin of the Seismological Society of America, 2005, 96(4B):S206-S220.

[39] 张肇诚, 王贵宣. 地震前兆含义,科学问题与研究途径的研讨[J]. 地震, 1997(4):429-439.

[40] 吴忠良，王林瑛. 可能的地震前兆的一个统计性质及其与地震类型的关系[J]. 地震学报, 2004(S1):59-64+177.

[41] 李献智. 应用地震事件判定前兆异常的可靠性[J]. 地震研究, 1996(2):121-126.

[42] Zhu Y Q, Liang W F, Zhang S. Earthquake precursors: spatial-temporal gravity changes before the great earthquakes in the Sichuan-Yunnan area[J]. Journal of Seismology, 2018, 22(4):1-11.

[43] 陈学忠，王晓青，李志雄. 强震前水氡异常台站时空分布非均匀性变化特征[J]. 地震，2000, 20(1):39-44.

[44] 郝建国，潘怀文，毛国敏等. 准静电场异常与地震——一种可靠短临地震前兆信息探索[J]. 地震地磁观测与研究, 2000, 21(4):3-166.

[45] 徐世浙. 评价地震预报效果的一种方法[J]. 地震, 1982(5):15-15.

[46] 三性法与静中动判据预测大地震的应用研究及物理基础探讨[D]. 中国地震局兰州地震研究所, 2012.

[47] 荣玉仿，David，Jackson D . 地震概率预报和检验的方法及应用[C]// 中国地球物理学会年刊2002——中国地球物理学会第十八届年会论文集. 2002:234-235.

[48] Vahaplar A, Tezel B T, Nasiboglu R, et al. A monitoring system to prepare machine learning data sets for earthquake prediction based on seismic-acoustic signals[C]// International Conference on Application of Information & Communication Technologies. 2015.

[49] Asim K M, Martínez-Álvarez F, Basit A, et al. Earthquake magnitude prediction in Hindukush region using machine learning techniques[J]. Natural Hazards, 2016, 85(1):1-16.

[50] Yue L, Yuan L, Li G, et al. Constructive Ensemble of RBF Neural Networks and Its Application to Earthquake Prediction.[C]// International Conference on Advances in Neural Networks. 2005.

[51] Morales-Esteban A , F. Martínez-álvarez, Troncoso A , et al. Pattern recognition to forecast seismic time series[J]. Expert Systems with Applications, 2010, 37(12):8333-8342.

[52] 朱海宁. 基于改进支持向量机回归的地震预测方法研究[D]. 2016.

[53] Devries P M R, Viégas F, Wattenberg M, et al. Deep learning of aftershock patterns following large earthquakes[J]. Nature, 2018, 560(7720): 632-634.

[54] Ding Y, Peng Y, Jie L I. Cluster Analysis of Earthquake Ground-Motion Records and Characteristic Period of Seismic Response Spectrum[J]. Journal of Earthquake Engineering, 2018(1):1-22.

[55] Florido E, Martínez-Álvarez F, Morales-Esteban A, et al. Detecting precursory patterns to enhance earthquake prediction in Chile[J]. Computers & Geosciences, 2015, 76(C):112-120.

[56] Stiros S C, Papageorgiou S. Seismicity of Western Crete and the destruction of the town of Kisamos at AD 365: Archaeological evidence[J]. Journal of Seismology, 2001, 5(3):381-397.

[57] Ikram A, Qamar U. Developing an expert system based on association rules and predicate logic for earthquake prediction[J]. Knowledge-Based Systems, 2015, 75(C):87-103.

[58] Zhang J, Lu Y L, Wu S C. Earthquake Prediction Research Based on the Mining in Time Series of Groundwater Temperature[J]. Advanced Materials Research, 2012, 532-533:1016-1020.

[59] Mirrashid M. Earthquake magnitude prediction by adaptive neuro-fuzzy inference system (ANFIS) based on fuzzy C-means algorithm[J]. Natural Hazards, 2014, 74(3):1577-1593.

[60] 晏昱. 基于数据挖掘的短临地震预测[D]. 南京航空航天大学.

[61] 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, v.54；No.287(03):32-39.

[62] 金秀如, 雍珊珊, 王新安等. 地震监测系统 AETA 的数据处理设计与实现[J]. 计算机技术与发展, 2018, v.28;No.249(01):51-56.

[63] 曾敬武, 雍珊珊, 郑文先等. 适用于大地震临震预测的地声传感单元[J]. 计算机技术与发展, 2015, 25(12):133-137.

[64] 庞瑞涛 , 雍珊珊 , 王新安等. 地震监测系统的电磁信号的采集设计与实现[J]. 计算机技术与发展, 2018,28(02):27-30.

[65] 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(4):54-62.

[66] 雍珊珊, 王新安, 庞瑞涛等. 多分量地震监测系统 AETA 的感应式磁传感器磁棒研制[J]. 北京大学学报(自然科学版), 2018, v.54；No.287(03):40-46.

[67] Uyeda S, Hayakawa M, Nagao T, et al. Electric and magnetic phenomena observed before the volcano-seismic activity in 2000 in the Izu Island Region, Japan[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(11):7352-7355.

[68] Hattori K, Takahashi I, Yoshino C, et al. ULF geomagnetic field measurements in Japan and some recent results associated with Iwateken Nairiku Hokubu earthquake in 1998[J]. Physics & Chemistry of the Earth, 2004, 29(4):481-494.

[69] Di Toro G , Han R , Hirose T , et al. Fault lubrication during earthquakes[J]. NATURE, 2011, 471(7339):494-498.

[70] 郝锦琦, 顾芷娟. 实验室中与地质年代中岩石形变引起的磁化率各向异性[C]// 全国古地磁学、环境磁学与岩石磁学学术会议. 1997.

[71] 郭自强, 周大庄, 马福胜,等. 岩石破裂中的电子发射[J]. 科学通报, 1987, 32(11):879-879.

[72] Molchanov O A, Kopytenko Y A, Voronov P M, et al. Results of ULF magnetic field measurements near the epicenters of the Spitak (Ms = 6.9) and Loma Prieta (Ms = 7.1) earthquakes: Comparative analysis[J]. Geophysical Research Letters, 2013, 19(14):1495-1498.

[73] 中国地震局. 地震群测群防工作指南[M]. 地震出版社, 2004.

[74] Kumar S , Vig R , Kapur P . Development of Earthquake Event Detection Technique Based on STA/LTA Algorithm for Seismic Alert System[J]. Journal of the Geological Society of India, 2018, 92(6):679-686.

[75] Hallac, D., Vare, S., Boyd, S., & Leskovec, J. (2017). Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. KDD, 2017, 215-223. doi:10.1145/3097983.3098060.

[76] Agrawal, S., Atluri, G., Karpatne, A., Haltom, W., Liess, S., Chatterjee, S., & Kumar, V. (2017). Tripoles. Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17.

[77] De La Cal E, Villar J R, Vergara P, et al. A SMOTE Extension for Balancing Multivariate Epilepsy-Related Time Series Datasets[C]. International Joint Conference SOCO’17-CISIS’17-ICEUTE’17 León, Spain, September 6–8, 2017, Proceeding, 2018: 439-448.

[78] Kavšek B, Lavrač N, Jovanoski V. APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery[M]// Advances in Intelligent Data Analysis V. 2003.

[79] Borgelt C. An implementation of the FP-growth algorithm[C]// 2005.

[80] Turney P. Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm[M]. 1994.

[81] Chang C C, Lin C J. LIBSVM: A library for support vector machines[M]. 2011.

[82] Gusain K, Gupta A, Popli B. Transition-Aware Human Activity Recognition Using eXtreme Gradient Boosted Decision Trees[M]. 2018.

[83] K. Holtzman B, Paté A, Paisley J, et al. Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field[M].2018

[84] 冯志生. 地磁日变化畸变地震预报指标研究进展[J]. 国际地震动态, 2018, No.476(08):24-25.

[85] 甘颖,梁意文,谭成予等. 基于危险理论的地震预测方法[J].计算机工程,2019,45(01):278-283.

[86] 姚丽. 地磁低点位移法应用进展[C]// 中国地震学会地震电磁技术专业委员会地震电磁新技术新方法研讨活动论文摘要集. 2016.