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

地震灾害对人类的生命和财产造成了巨大的损失，准确地预测地震发生的时间、地点和震级可极大地减少损失。然而，现有地震预报工作仍处在探索阶段，地震的预测工作依然是待解决的难题。地震预测研究还需发展更多的观测方法，提取更多震前的异常信息。因此，北京大学深圳地震监测预测技术研究中心自主研制了多分量地震监测系统 AETA。目前系统已经稳定运行 2 年，布设超过 200 个台站，如何利用这些数据进一步提取与地震相关的特征并对地震三要素进行预测，已经成为亟需解决的问题。基于以上，本文进行了基于 AETA 系统数据的特征权重评估及地震预测模型的研究。

本文完成的主要工作和创新点如下：

1.本文提出了基于 AETA 数据波形特点的特征提取方法。首先，基于电磁数据出现日周期的波形特点，提取出了 7 种波形，并对台站进行波形相似性分析，提取波形出现的次数，共得到 21 种特征。其次，基于地声在震前会出现尖峰的特点，对波峰进行提取，共得到 9 种特征。

2.本文基于信号变换方法对电磁数据和地声数据进行特征提取。本文提出了基于希尔伯特-黄变换的能量分析法。首先将电磁和地声信号利用 EMD 进行分解，选取 imf1 分量。之后，对其进行希尔伯特变换得到震前波形的瞬时能量变化，共得到 36 种特征。

3.本文对所提取的 66 种特征进行分析，发现其中包含一些冗余特征。这些冗余的特征会导致分类器计算开销太大，性能较差。因此，本文对 relief-F、LVW、随机森林三种代表性的特征选择算法进行了研究对比并最终利用随机森林对所提取的特征集进行了权重评估，最后筛选出了对地震预测贡献度最大 energy_peak_max、ring_15、energy_sum 等 20 个特征。

4.本文以地震时间、震级、震中三要素的预测为目标，利用筛选后的 20 个特征，使用梯度提升树建立了地震预测模型，并且在东经 100°~103°，北纬 25°~28°以及东经 103°~106°，北纬 31°~34°两个区域，2017 年 6 月 1 日至 2019 年 3 月 1 日期间的数据集上进行了数据实验，得到模型在区域 1 验证集中地震事件的准确率 0.68，查全率为 0.57；在区域 2 验证集中地震时间的准确率为 0.56，查全率为 0.62。区域 1 验证集中震中经纬度预测均方误差分别为 0.48 和 0.71。区域 2 验证集中震中经纬度均方误差分别为 1.31 和 0.80。实验结果表明，本文的工作对于地震预测有一定意义。

关键词：地震预测，特征选择，随机森林，梯度提升树，希尔伯特-黄变换

Research Of Feature Weight Assessment and Earthquake Prediction Model Based on AETA Data

Jiyan Zhang (Microelectronics and Solid-State Electronics)
Directed by Xin'an Wang

ABSTRACT

Earthquake disasters have caused tremendous losses to human life and property. Accurate prediction of the time, location and magnitude of earthquakes can greatly reduce losses. However, the current work of earthquake prediction is still in the exploratory stage, and the work of earthquake prediction is still a difficult problem to be solved. In the study of earthquake prediction, more observation methods need to be developed to extract more anomalous information before earthquakes. Therefore, the Shenzhen Seismic Monitoring and Prediction Technology Research Center of Peking University independently developed the multi-component seismic monitoring system AETA. At present, the system has been running steadily for two years, with more than 200 stations installed. How to use these data to further extract the characteristics related to earthquakes and predict the three elements of earthquakes has become an urgent problem to be solved. Based on the above, this paper carries out the research of feature weight evaluation and earthquake prediction model based on the data of AETA system.

The main work and innovation of this paper are as follows:

1. Based on waveform characteristics, this paper extracts features from electromagnetic data and geoacoustic data. Based on the waveform characteristics of the daily period of electromagnetic data, seven waveforms are extracted, and the similarity of waveforms is analyzed at the stations. The number of waveforms is extracted and 21 features are obtained. Based on the characteristics that the ground acoustic peak appears before the earthquake, nine features are obtained by extracting the wave peak.

2. This paper extracts the features of electromagnetic data and geoacoustic data based on signal transformation. In this paper, an energy analysis method based on Hilbert-Huang transform is proposed. Firstly, electromagnetic and acoustic signals are decomposed by EMD, imf1 components are selected, and the instantaneous energy changes of waveforms before earthquakes are obtained by Hilbert transform. A total of 36 characteristics are obtained.

3. Based on the analysis of 66 features extracted, some redundant features are found. These redundant features will result in too much computational overhead and poor performance of the classifier. Therefore, three representative feature selection algorithms, relief-F, LVW and random forest, are studied and compared in this paper. Finally, the weights of the extracted feature sets are evaluated by using random forest and screen out 20 features, such as energy_peak_max, ring_15 and energy_sum, which have the greatest contribution to earthquake prediction.

4. In this paper, the prediction of earthquake time, magnitude and epicenter is taken as the target, and 20 selected features are used to establish the earthquake prediction model using gradient lifting tree. Data experiments are carried out on the data sets of the period from June 1, 2017 to March 1, 2019 in the two regions of 100 to 103 degrees in the East longitude, 25 to 28 degrees in the North latitude, 103 to 106 degrees in the east longitude and 31 to 34 degrees in the North latitude. The accuracy of area 1 is 0.68, recall is 0.57, and that of area 2 is 0.56, recall is 0.62. The mean square errors of longitude and latitude prediction of centralized earthquakes verified by region 1 are 0.48 and 0.71, respectively. The mean square errors of longitude and latitude of the centralized epicenter are 1.31 and 0.80 respectively. The experimental results show that the work in this paper has certain significance for earthquake prediction.

KEY WORDS: Earthquake Prediction, Feature Selection, Random Forest, Gradient lifting tree, Hilbert-Huang Transform

目录

第一章　绪论

1.1　研究背景

地震是对人类社会危害最大的自然灾害之一。近年来，全球曾发生多次强震，如：2008 年中国汶川发生的 8.0 级地震、2009 年印度尼西亚苏门答腊岛南部发生的 7.7 级地震、2011 年日本本州岛东部发生的 9.0 级特大地震、2017 年中国九寨沟发生的 7.0 级地震等。我国是一个地震频发的国家，据统计，自 21 世纪至今，已发生 300 余次 5 级以上的地震，平均每年约发生 20 余次地震。其中一些地震造成了大量的人员伤亡和难以估量的财产损失。如果发震之前能够准确预测地震三要素，就有机会降低地震对人类社会带来的损失[1]。

地震预测是一个极具挑战性的科学难题，在很早以前就是学术关注的热点，但是由于地震的复杂性，现在的地震预报仍然是低水平的探索阶段[2]。从根本上说，其主要原因：一是至今有限的科学知识水平，单纯地从理论上还没有办法能够十分清楚地认识地震孕育、发生的规律；其二是现有的地震活动的监测技术还没有达到普适的效果[3]。目前，国内外对此进行大量的研究，表明可观测地震前兆以实现地震的预测和预报工作，目前主要进行的震前前兆观测的手段有地磁以及地电流观测[4-7]、次声波异常观测[8,9]、地壳变动的连续观测[10,11]、电流层扰动观测[12,13]、卫星红外异常观测[14-16]、地下水水位和化学元素含量的观测[17,18]等。

多分量地震监测预测系统 AETA（简称 AETA）正是在前兆观测进行地震预测的需求下应运而生。AETA 系统由于具有低成本、覆盖范围广、密度大、实时监测等优点，为动态监测地震的孕育、发生过程提供了一种可能，为地震预测的前兆监测方法提供了一种新的技术手段。本文基于北京大学地震预测技术研究中心自主研发的 AETA 系统所采集的电磁和地声数据进行震前的特征提取，以及地震预测模型研究。

1.2　国内外研究现状

1.2.1　地震预测方法的研究状况

地震预测的复杂性及困难性是世界公认的，从 19 世纪中期开始，学者们逐步开展了各种手段的地震预测研究。其中，地震学家 Milne[19]于 1880 年便开始对地震预测这

一科学难题进行探究。Cornell[20]最早使用泊松分布模型来预测地震并提出为线性分析方法（PSHA）的基本理论框架。因为地震预测的复杂性，时至今日，地震预测仍然处在摸索的阶段，还没有形成一个成熟、完善的地震预测理论。但许多国家在地震活动的特征和规律，地震前兆观测、地震发生的机理等方面研究还是取得了一定的进展。

国内外学者对地震发震时间、震级、震中的预测研究主要有以下四种思路：

1.地震与地质有关。因为大多数的地震都发生在地壳的中上层，所以地震发生的过程应该属于地质发生变化的过程，若对地质构造特征进行研究将对地震三要素的预测有所贡献。

Papazachos 等人[21]发现在强主震发生之前，距震中较远的区域内发生了中震级的前震产生的地震应力加速改变以及距震中较近的区域发生了小震级的前震发生的地震应力减速现象。在此地震活动模式的基础上，建立了中期地震预测的模型。

尹祥础[22]在 1987 年提出利用加卸载响应比的方法可进行中短期地震预测。此方法表明地震发生的过程是在一个非线性、极其复杂的过程。地震时由于应力导致地震中心区域的内部介质发生失稳和破裂，因此如果系统出现加卸载响应的比值异常变动较大，则可以判断地震发生的危险程度。张浪平[23]通过对加卸载响应比异常区域演化过程的研究，进行判断伊朗地区未来的发震情况。秦四清[24,25]等在 2010 年发现孕震断层中每个锁固段断裂点与加速应变能释放起点的累积 Benioff 应变之比，可能为锁固段个数底数为 1.48 的指数关系，初步认为 1.48 很可能是中等强度及其以上震级地震产生过程的普适常数。此方法可适用于地震的中期、短、临震预报。

2.运用统计机器学习相关的方法进行地震预测。根据历史上的统计出的震例，可以从这些震例所含的参数之中总结出地震事件序列的规律，从而对地震进行预测。

我国学者陈祺福[26]在 1997 年第一次创建了基于遗传算法的地震预报分类体系，并用此方法对“首都圈”的地震预测进行了初步的研究。后来，王海涛等[27]设计了基于遗传算法的短期地震综合预报的方法，李莹甄[28]用此方法对北天山地震带进行了地震预测，并取得了不错的预测效果。李荣峰[29]在 2000 年曾经基于神经网络模型对福建以及其周边区域进行年度最大震级预测，预测准确率达 90%。

Rundle[30]等在 2000 年将旋节线的相分离与地震发生前的活动性行为的关系进行探究。Holliday[31]等在 2005 年使用图像信息法研究之后的十年内全球 7 级以上地震，结果显示，2014 年 12 月 23 日 8.1 级地震以及 2004 年 12 月 26 日的 9 级地震，两次大震均在地震热点或在其周围。Asim K M 等[32]在 2018 年通过利用最大关联度和最小冗余度（MRMR）标准提取相关特征，建立了基于支持向量回归器（SVR）和混合神经网络（HNN）的地震预报分类系统，并应用于兴都库什、智利和南加州地区。在与以往的预测研究相比，预测性能都有所改善。日本学者 Nanjo[33]在 2006 年使用图像信息法对同一时期内在日本发生的 Niigata6.8 级地震做出了较好的预测。

3.基于前兆研究的地震预测。基于前兆研究的主要思路是认为地震过程属于物理变化过程，地震发生前的前兆信号的异常变化，并研究这些信号与地震的关系，从而预测地震的发生，比如地下水变化，电离层变化等异常现象。

张炜等[34]在 1987 年对观测地下水或土壤中气体的化学组分变化进行研究，并根据监测氡的含量变化进行预报地震的研究工作。张桂清等[35]研究了全球地震活动与太阳活动之间的关系，研究结果表明地震活动性和太阳活动存在着周期影响。

国内外学者一直在研究地震和地磁场的关系，震磁效应已经被广泛地应用在地震的短临预报中[36]。Zeng Xiaoping[37]在 2001 年对在 1977 年发生的巴尔干 6.4 级地震前后地磁数据主成分进行分析，发现其变化的比率可作为地震发生的判断依据。Suratgar 等[38]在 2008 年通过神经网络方法利用地磁场偏角、水平分量等变化进行地震发生前的 2 天的震级预测，得到良好的效果。

电磁扰动一直作为前兆观测的方法，可捕捉在地震发生前的异常信号。Masashi Hayakawa 等[39]认为岩石圈超低频电磁辐射和地震电离层扰动可用来进行短期地震预测。中国在 20 世纪中期便开始进行电磁扰动的研究。迄今为止，经过许多地震研究学者的潜心努力，已经有所发现[40-43]。其中，丁跃军[44]等分析了在 2008 年汶川 8 级地震发生前的 1 个月，空间电离层和地面电磁均出现多次明显的异常。Kopytenko 等[45]在 2001-2003 年期间，进行定期观测超低频电磁扰动。在此期间，发现 5 个 5 级以上地震发生前 12-18 小时观测到异常强度增加。

4.通过多方法组合预测地震。此方法主要是指将多种前兆方法组合联系组合进行预测地震。

国内外学者曾对此进行了深入的研究与探讨[46,47]。Bowman 和 King[48]把库仑应力和矩率加速变化相结合，并对自 1950 年开始加利福尼亚所有的 6.5 级以上地震进行验证，发现地震前有明显的加速过程。Gelfand[49]在 1976 年利用模式识别的方法将几种地质特征进行组合，并利用迭代的方法进行地震预测。。从 2009 年 7 月到 2011 年 12 月(19 个目标地震),利用迭代的方式进行组合模型具有更好的预测性能。许多目标地区的地震发生的组合模型具有较高的预测率,这些地区的预测准确率远高于简单的平均模型的预测准确率。Shebalin 等[50,51]提出了基于差分概率增益的迭代方法将多个模型进行组合，此组合方法预测的准确率比单个方法高。薄万举[52]提出将多个单项地震前兆信息进行信息合成，此方法显著地优化了综合多种前兆信息进行预测地震的效果。韩天锡等[53]提出了基于判别分析、主成分分析的多元统计组合，并在华北地区进行地震综合预报，取得了良好的效果。王海涛[54]基于综合异常指数提取了多种前兆特征的综合群体特征，并对新疆乌苏地震进行研究，结果表明，两组地震发生前的综合异常指数均出现了明显的高值异常变化过程，说明在此方法能够表示地震前兆异常的综合变化。

1.2.2 地震前兆观测的研究状况

随着国内外学者在地震领域研究的深入，通过地震发生前的异常记录得到的地震前兆信号也越来越多，其中在地震发生前震中以及附近区域出现的异常变化，如：地声变化、电磁场变化、应力场变化、地下水等都可以作为震前出现的前兆特征。

地壳形变和重力异常在大地震发生之前出现了明显的异常。其中，在 1996 年，发现了地震与出现的一些大面积的地壳形变速率存在一定的关系[55]；2005 年，楼海和王椿镛通过小波分解，也发现川滇地区出现明显的重力异常[56]。

在水文和地球化学方面，在大地震发生之前地下水的化学成分会出现明显的数值异常，其中，王吉易在 1988 年发现地震前的水氡出现明显的异常[57]，赵永红等在 2011 年也发现氢同位素在震前也出现明显的异常[58]。

在地磁观测方面，针对地磁异常的观测和研究一直都是国内外学者的研究热点。苏联最先进行地磁观测，随后日本、美国等各国家也逐步开始进行地震前兆地磁异常方面的观测和研究。美国斯坦福大学通过 0.01-10Hz 的观测频段研究发现，震前地磁异常变化的幅值受很多因素的影响，但在大多数的情况下，异常主要出现在距离地震发生前的 1~19 天内，而且会持续几分钟到 1 天之内[59]。

自 21 世纪初我国组建了属于自己的数字地震台网，中国地震研究学者便利用监测的数据进行异常信号的识别。2001 年 11 月 14 日在昆仑山发生的 8.1 级地震，在分析地震计在地震发生前监测的数据之后，结果发现在距离震源中心 700 公里以外的地震计监测到了明显的前驱波，波形与纺锤相类似[60]。洪星[61]在 2004 利用数字地震计监测到的 sPn 震相，得到根据多台地震计监测的 sPn 震相，可用来进行预测地震的震源深度。

从 20 世纪中期开始，国内外学者便开始进行电离层的观测和研究。其中，Barnes 和 Leonard 在 1965 年发现在阿拉斯加大地震期间出现明显的电离层异常扰动[62]。Parrot[63]对由于地球自然物理活动造成的电磁扰动与地震之间的关系进行了探究。Pulinets[64]则对强震发生前的几天或几小时内电磁层的变化进行了探究，并对此建立了物理机制。

地声作为一种地震震前前兆特征，国内外学者已进行了研究与探讨,以期待实现地震预测预报工作。早在 20 世纪初就已经有地声的观测记录，之后美国、日本等国家相继对地声进行了深入的研究[65]。而我国从 20 世纪中期开始进行地声的观测与研究，并且在 20 世纪 80、90 年代达到高潮。

自从前苏联科学家 Grony 等[66]在 1988 年利用 NOAA 卫星采集的亮度温度数据，发现在中亚地区的一些中强地震发生之前出现热红外异常现象。热红外异常便开始作为地震前兆异常的特征，引起国内外的的学者开始陆续进行卫星红外数据以及其反演

的参量热辐射异常的研究，这些参量主要有：亮度温度 、长波辐射、地表温度、潜热通量等[67-70]。

1.2.3　地震前兆特征提取方法与基于前兆特征的预测方法

目前，有些地震前兆特征无法直接获取，需要通过一定的研究方法进行提取。现如今，主要利用信号变换、数理统计、机器学习等方法进一步提取前兆特征可能隐含有用的信息，利用这些特征进行地震预测[71,72]。

张轶鹏[73]利用形态学区域生长算法对地震发生前的电磁异常进行提取，并对 2006 年所发生的地震进行异常检测，取得了较好的效果。周挚等[74]基于 HHT 时频分析应用在重力固体潮前兆特征提取，并在昆明、下关进行实验，所提取的特征在震前出现同步的异常现象。张璇等[75]在 2018 年利用 7 阶小波插值法对红外亮温资料进行分析，发现在汶川地震前出现两条突出的红热外条带，并且随着距离发震的时间越近，异常强度越大，在发震前来两天达最大值。

王乾龙等[76]利用长短期记忆网络(LSTM)的深度学习技术，研究不同地点地震的时空关系，并利用这种进行地震预测。实验结果表明，此方法能够找到地震之间的时空相关性，可以更好地进行预测地震活动。

Mostafa Allamehzadeh[77]在 2017 年利用 Kohonen 自组织神经网络预测地震发生的位置。并在伊朗的阿尔伯茨地区进行了实验，结果表明，利用此方法预测了 5.5 级地震发生的位置。并提出利用 MZ 方法对主要断裂带进行识别，并在此基础上利用模式识别的方法生成自组织特征图确定危险的区域。

Asim 等[78]在 2018 年利用古登堡-里氏定律、地震速率变化、地震能量释放等方法提取了 60 种地震特征，并根据最大关联度和最小冗余度标准提取地震前兆特征。依此特征建立了基于支持向量回顾和混合神经网络的地震预报分类系统。此方法应用在兴都库什、智利和南加州地区，结果表明，利用此方法进行地震预测取得了良好的效果。

1.3　主要研究内容

本文的数据来源于北京大学深圳研究生院地震监测预测技术研究中心自主研发的设备——多分量地震监测预测系统 AETA 所监测的地声数据和电磁数据。根据国内外进行地震预测方法的相关研究，为了可以进一步根据 AETA 数据进行地震的预测研究，本文主要做了以下几方面的研究：

1.根据 AETA 数据的特点进行数据预处理

本文主要针对本文所研究的地震数据集所具有的特点，采用相应的方法，进行缺失数据的补充，对存在干扰信号的数据进行处理，之后对数据进行归一化处理，使得到的数据满足数据挖掘、特征提取的需求。

2.基于 AETA 数据进行特征提取

本文提出了针对 AETA 数据波形的特点进行特征提取的方法。首先根据电磁数据的特点，发现电磁数据在震前会出现几种特征波形，因此本文对震前出现的”SRSS”波及类“SRSS”波形进行编码，利用波形相似性分析，对震前 15,10,7 天的数据进行波形的特征提取，共得到 21 种特征。然后基于地声数据在震前出现尖峰的特点，利用欧式距离对尖峰进行提取，共得到 9 种特征。最后基于希尔伯特-黄变换对电磁数据和地声数据进行瞬时能量分析，得到电磁数据以及地声数据震前 15,10,7 天的瞬时能量变化，共得到 36 种特征。

3.特征权重评估

首先对 relief-F 算法，LVW 算法，以及随机森林进行研究分析并对比其优缺点。最后基于第三章提取的特征利用随机森林研究评估分类过程中各特征的重要性，并筛选出了对地震预测贡献度最大 energy_peak_max、ring_15、energy_sum 等 20 个特征。

4.地震预测模型研究

利用之前所筛选后的特征利用 cart 决策树，支持向量机，梯度提升树进行地震震级预测的研究。利用筛选后的 20 个特征，使用梯度提升树建立了地震预测模型，并且在东经 100°~103°，北纬 25°~28°以及东经 103°~106°，北纬 31°~34°两个区域，2017 年 6 月 1 日至 2019 年 3 月 1 日期间的数据集上进行了数据实验，得到模型在区域 1 验证集中地震事件的准确率 0.68，查全率为 0.57；在区域 2 验证集中地震时间的准确率为 0.56，查全率为 0.62。区域 1 验证集中震中经纬度预测均方误差分别为 0.48 和 0.71。区域 2 验证集中震中经纬度均方误差分别为 1.31 和 0.80。

1.4 论文组织结构

本文利用多分量地震监测预测系统 AETA 的数据进行特征提取，特征选择，以及地震预测模型建立，以地震预测模型为核心进行论述。全文总共为六章：

第一章：绪论。首先，研究了地震预测的研究背景和意义。然后，从地震预测方法、前兆观测、以及前兆特征提取三个方面的国内外现状进行分析。最后介绍了本文的研究内容和研究方法，以及论文组织结构。

第二章：AETA 数据分析与处理。首先对本文数据来源多分量地震监测系统 AETA、地声探头、电磁探头进行阐述。然后基于震例对电磁、地声数据进行分析。最后基于

AETA 数据的特点，对缺失值进行填充，对由于断电导致的非正常信号进行处理，并对处理后的信号进行归一化处理。

第三章：基于 AETA 数据的特征工程。首先根据电磁数据的特点，发现电磁数据在震前会出现几种特征波形，因此本文对震前出现的”SRSS”波及类“SRSS”波形进行编码，利用波形相似性分析，对震前 15,10,7 天的数据进行波形的特征提取，共得到 21 种特征。然后基于地声数据在震前出现尖峰的特点，利用欧式距离对尖峰进行提取共得到 9 种特征。最后基于希尔伯特-黄变换对电磁数据和地声数据进行瞬时能量分析，得到电磁数据以及地声数据震前 15,10,7 天的瞬时能量变化，共得到 36 种特征。

第四章：基于震例的特征权重评估。首先提出了特征选择的原因以及一般流程，并根据不同的评价函数的过滤式、包裹式、嵌入式特征选择算法进行了原理及优缺点进行了分析。然后研究对比了 relief-F 算法，LVW 算法，以及随机森林算法在特征选择中的表现并最后选择随机森林对第三章所提取的特征进行评估，计算得到特征的重要性，并筛选出对地震预测贡献度最大 energy_peak_max、ring_15、energy_sum 等 20 个特征。

第五章：地震预测模型的研究。首先提出了地震三要素的预测结构，研究了地震三要素的选择。然后对比了 cart 决策树，支持向量机，梯度提升树的地震震级的预测效果。本文提出了基于 AETA 数据的地震地点与震级预测的方法。利用筛选后的 20 个特征，使用梯度提升树建立了地震预测模型，并且在东经 100°~103°，北纬 25°~28°以及东经 103°~106°，北纬 31°~34°两个区域，2017 年 6 月 1 日至 2019 年 3 月 1 日期间的数据集上进行了数据实验，得到模型在区域 1 验证集中地震事件的准确率 0.68，查全率为 0.57；在区域 2 验证集中地震时间的准确率为 0.56，查全率为 0.62。区域 1 验证集中震中经纬度预测均方误差分别为 0.48 和 0.71。区域 2 验证集中震中经纬度均方误差分别为 1.31 和 0.80。

第六章：总结与展望。总结了本文所完成的工作以及最终的结果，并对未来的改进和发展方向进行了展望。

第二章　AETA 数据处理与分析

2.1　数据来源

2.1.1　AETA 系统介绍

从 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便开始研发多分量地震监测系统 AETA[79](简称 AETA)，由数据处理终端[80]、地声传感探头[81]、电磁传感探头[82]以及监测数据云平台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[83,84]。地声和电磁传感探头可置于山洞内或埋在浅表 2m 以下的位置，数据处理终端置于机房。AETA 系统降低了安装的时间和要求，使 AETA 系统可大批量地进行安装布设。

AETA 系统分为数据采集与数据分析两个方面：

（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。

图 2.1　AETA 系统实物图

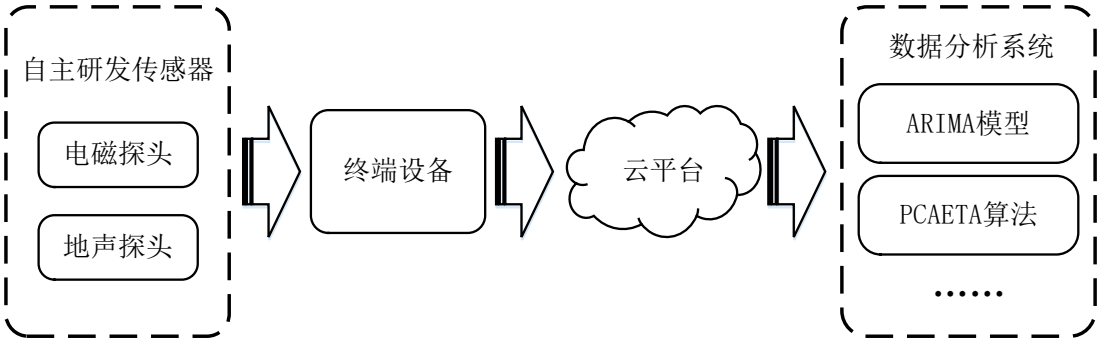


图 2.2　AETA 系统框图

AETA 系统的数据主要由终端经过处理传到云平台进行存储。数据的显示部分可通过网页查看，如图 2.3 所示。数据分析系统的主要功能是把 AETA 监测的数据转换成特征数据，主要分为均值、振铃计数、峰值频率三种分量。

图 2.3　AETA 数据查看网页

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，两版设备均进行了现场试验。结果显示，AETA 系统对当地所发生的地震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，多分量地震监测预测系统 AETA 正式定型批量生产。

截至目前，在国家地震局的支持下，AETA 系统已在全国范围内安装 200 余台，遍布河北、四川、云南、西藏、广东和台湾地区。图 2.4 表示的是多分量地震监测系统 AETA 的布设情况。

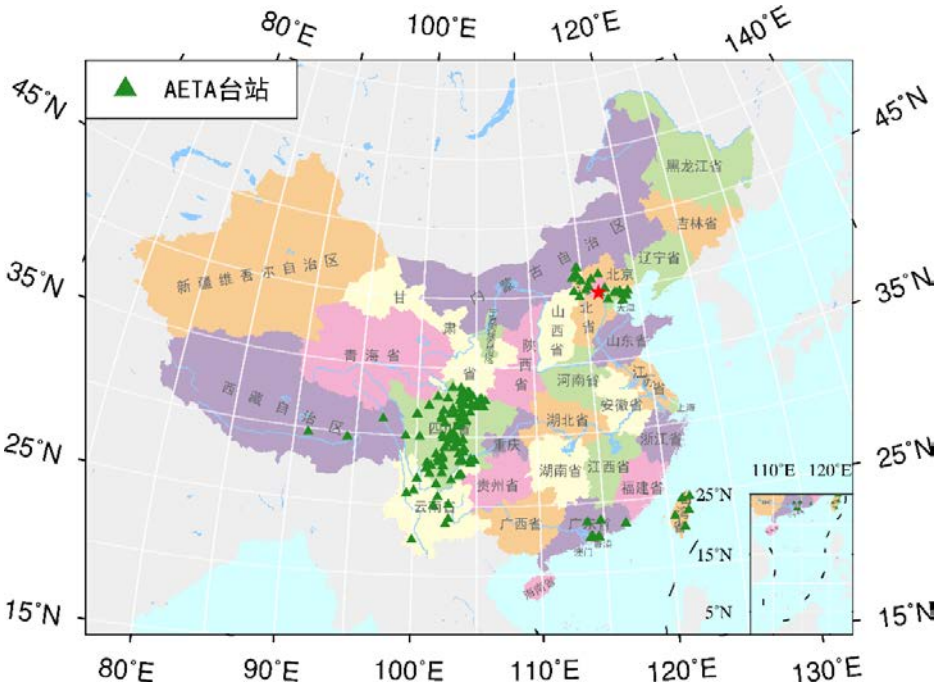


图 2.4　AETA 布设情况

2.1.2　电磁数据来源

本文采用的电磁数据来源于 AETA 电磁信号实时采集系统，如图 2.5 所示，包括信号调理和实时采集两部分。

图 2.5　电磁探头实物图

电磁信号采集系统包含：原始数据信号调理部分、数据采集部分及远程升级和参数配置部分。其中，原始数据信号调理部分首先将监测到的 0.1Hz-10kHz 的电磁信号进

行放大、磁反馈电路、滤波、次级放大、A/D 转换。然后将调理后的信号传给数据采集部分。其中，数据采集部分基于 STM32F407 与 W5300 的交互，首先将监测到的数据进行打包并发送到远程数据中心，之后与远程数据中心对数据进行后续的处理，可实现远程实时监控设备的运行情况[75]。

电磁传感探头：监测 0.1Hz-10kHz，0.1-1000nT 波段，采样率为低频 500Hz，全频 30kHz[73]。

具体参数如表 2.1 所示：

表 2.1　电磁数据参数表

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

2.1.3　地声数据来源

本文采用的地声数据来源于 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电路结构。

图 2.6　地声探头实物图

地声监测传感单元是通过将捕获的地声信号传导至压电薄膜传感器，将地声信号转化为电信号。首先预处理电路对信号进行处理，之后将预处理后的信号通过电缆传输至地面基站进行处理和发送。为了可以包含震前地声信号所涵盖的次声波、可听波、超声波范围，AETA 系统采用 PVDF 型薄膜传感器。其传感器的特性是：灵敏度为-180dB·V×104/Pa，频率响应范围为 10^{-3}Hz-1MHz，范围为-50～100℃，满足震前地声检测的要求。

地声传感探头参数：监测 0.1Hz-50kHz 波段，采样率位低频 500Hz，全频 150kHz[82]。具体参数如表 2.2 所示：

表 2.2　地声数据参数表

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

2.2　基于震例的数据分析

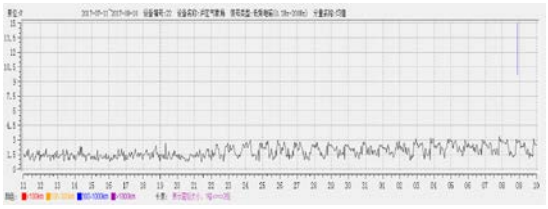
AETA 系统自 2015 年开始布设第一台，至今 4 年已布设超过 200 台，遍布四川，云南，西藏，广东，首都圈，台湾，孟加拉。至今已有多次映震效果，本文将以 2017 年 8 月 8 日 21:19:46 发生的 7.0 级地震作为典例，进行 AETA 电磁数据、地声数据的有效性分析。

2.2.1　基于震例的电磁数据分析

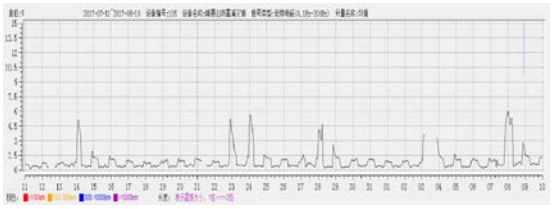
地震发生前，在四川布设的 AETA3.0 设备共计 36 台，不仅出现电磁临震异常，同时发现了可能与地震孕育过程有关的 SRSS 波现象。SRSS 波是指一种近似日周期的波形，该波形与日升日落几乎同步变化，日升时变低、日落时变高。AETA 出现电磁临震出现 SRSS 的台站有 6 个。

出现 SRSS 波的台站： 泸定气象局 LD、峨眉山防震减灾局 EMS、青川县防震减灾局 QC、茂县测点 MX、犍为防震减灾局 QW、西昌小庙山洞 XCXM。

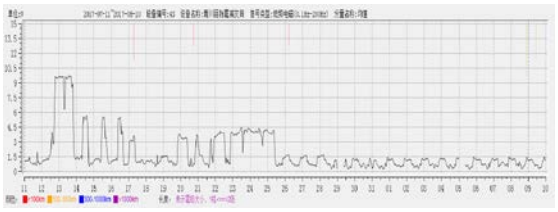
下面将相关台站的数据展示和说明如下：



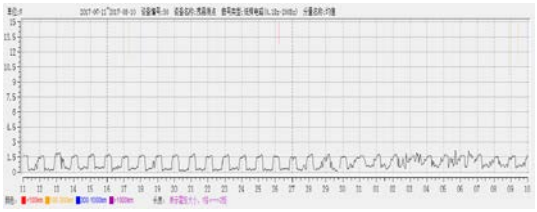
(a)泸定气象局



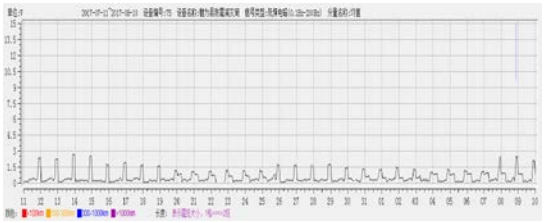
(b) 峨眉山防震减灾局



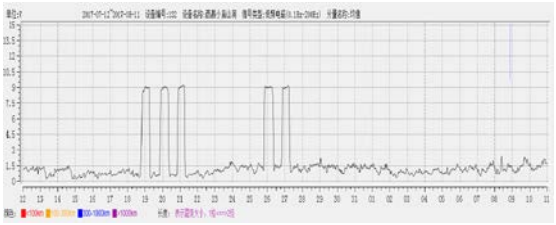
(b) 青川县防震减灾局



(d) 茂县测点



(e) 犍为防震减灾局



(f) 西昌小庙山洞

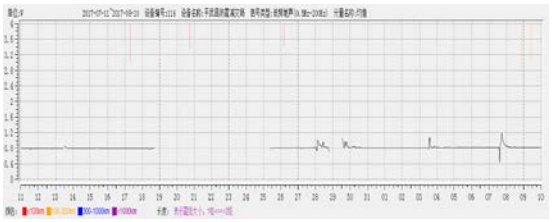
图 2.7　震前出现 SRSS 波台站波形图

从图中可以看出，（a）泸定气象局在震前 15 天开始出现规律的 SRSS 波；（b）峨眉山在震前也出现规律的 SRSS 波，并且在地震当天突然上升；（c）青川县防震减灾局在震前 13 天开始规律的 SRSS 波；（d）茂县测点在震前出现 SRSS 波，并在震前 7 天开始规律的 SRSS 波被打乱；（e）犍为防震减灾局在震前出现规律的 SRSS 波；（f）西昌小庙山洞在 2017 年 7 月 19 日至 7 月 21 日，7 月 25 日至 7 月 27 日出现 SRSS 波。

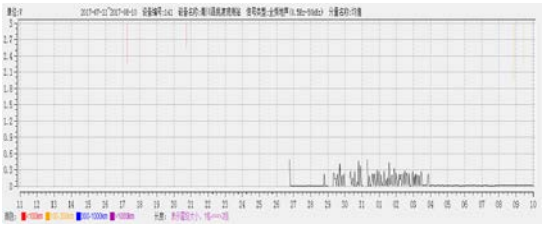
2.2.2　基于震例的地声数据分析

地震发生前，在四川布设的 AETA3.0 设备共计 36 台，AETA 出现地声临震异常的台站有 16 个。

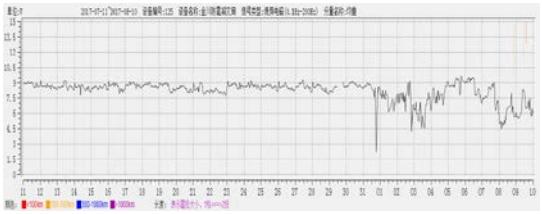
异常台站：平武防震减灾局 PW、青川姚渡观测站 QCYD、金川防震减灾局 JC、小金县防震减灾局 XJX、峨眉山防震减灾局 EMS、石棉挖角乡 SMWJ、马边地震局 MB。



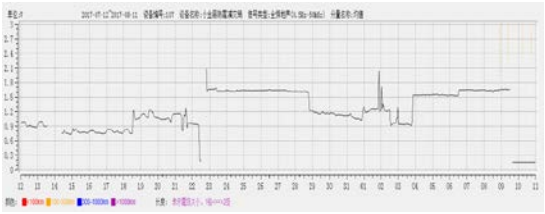
(a)平武县防震减灾局



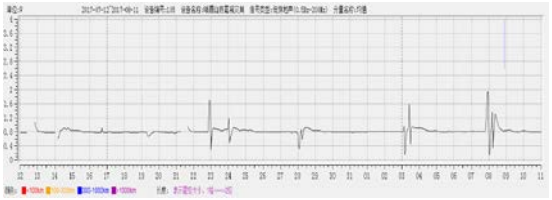
（b）青川姚渡观测站



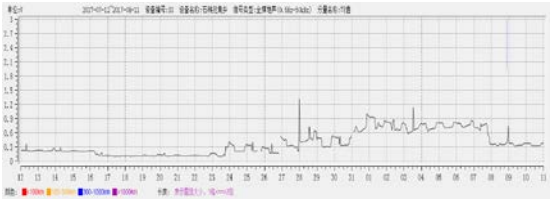
（c）金川防震减灾局



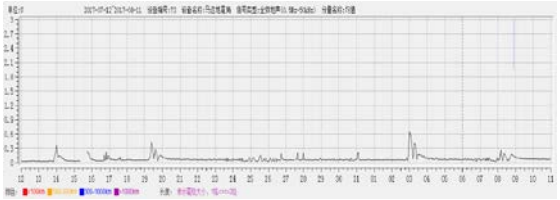
（d）小金县防震减灾局



（e）峨眉山防震减灾局



（f）石棉挖角乡



(g) 马边地震局

图 2.8　地声出现异常的台站波形图

从图中可以看出，地声的信号并非像电磁信号有剧烈的变化，但是会在地震发生之前出现明显的尖峰。

2.3　原始数据预处理

在实际数据采集过程中，不可避免地会产生由于断电断网而导致的数据缺失，为了保证数据的完整性和可用性，需要对数据进行必要的预处理。本章主要针对本文所

研究的地震数据集所存在的问题，进行缺失数据的补充，对存在干扰信号的数据进行处理。

2.3.1　数据集特点

本文所研究的数据来源于自主研制的 AETA 系统所提供的数据。数据中存在的主要问题是断网导致缺失数据以及断电导致的缺失数据。

1.断网导致的数据缺失

本文选取九寨沟防震减灾局 2018 年 1 月 10 日的数据为例，由于断电导致数据存在长度为几个小时的数据中断。

图 2.9　缺失数据波形图

2.断电导致的数据缺失

本文选取九寨沟防震减灾局 2018 年 1 月 10 日的数据为例，由于断电导致数据存在长度为几个小时的数据中断，当恢复电力供应时，数据出现由于温漂产生的无效数据。

图 2.10　异常信号波形图

2.3.2 断网导致的缺失数据处理

数据集有效数据的减少会致使数据挖掘的效果快速的下降，数据集若存在缺失的数据会对后续的数据挖掘结果产生较大的影响。

目前，国内外对缺失数据的处理主要有三种方法：删除、忽略、填充。为了能更好地进行后续的特征提取工作，本文采用线性插值法对缺失的数据进行填充缺失值。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，将采用线性插值法来对缺失的数据进行补全，如果连续缺失 10%以上的数据，插值补全会引入较大误差，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。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，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，具体算法如表 2.3 所示。

表 2.3　线性插值补全法

算法：线性差值补全法
输入：　时间序列: TS, 参数: l, fs, δ (typical value is 0.1).
输出：　一个完整的时间序列, CTS.
过程：
1: 初始化：　$\tau \leftarrow \emptyset$, $\gamma \leftarrow \emptyset$
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:　$\tau \leftarrow \tau \cup$ an array from t_i to t_{i+1} with step 1/ fs
5:　$\gamma \leftarrow \gamma \cup$ 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$\leftarrow(\tau, \gamma)$

对缺失数据补充后的数据如图 2.11 所示：

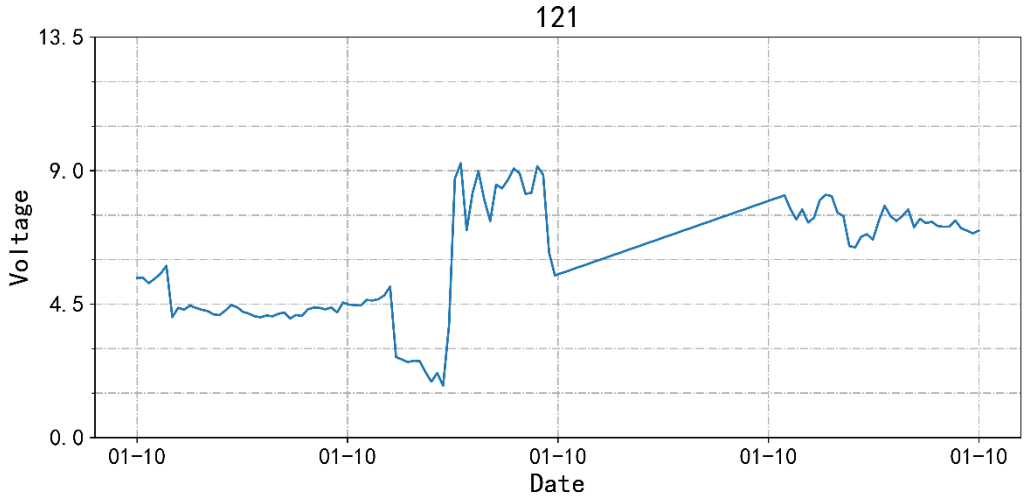


图 2.11　补全数据后的波形图

2.3.3　断电导致的的缺失数据处理

由 2.3.1 节可以看出由于断电，地声信号会出现一个尖峰，并会以指数型的下降的特点，本文采取的策略共分为以下五步：

1.利用数据库的日志查明数据中断的起始点；

2.设置阈值，若超过此阈值，则判断此时信号为数据异常信号为数据恢复的起点同时也为需要处理的异常信号的始端；

3.查找数据出现两次拐点的地方判断为异常信号的尾端；

4.将此段数据删除；

5.利用 2.3.2 节的数据填充的方法把数据进行填充。

得到的数据如图 2.12 所示：

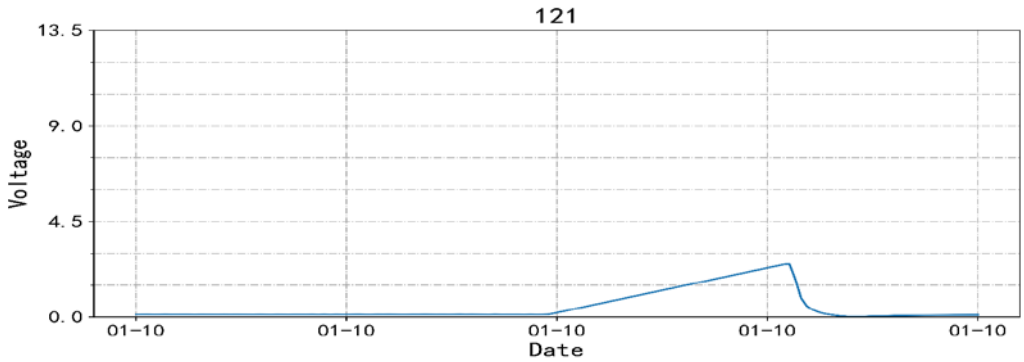


图 2.12　异常信号处理后波形图

2.3.4 数据归一化

数据的标准化，指的是首先选择一个区间，然后将样本数据利用相应的方法以一定的比例进行缩小或者放大，使最后的数据可以在之前规定的区间里。而归一化则是把规定的区间变为[0,1]。本文采用的数据来源于 AETA 在全国布设的台站实际监测的数据，但由于在实际的监测中，每个台站的数据的差异性较大。此种情况会导致无法提取有效的特征，因此需要对数据进行归一化。

归一化后的数据同时又具有以下优势：

1.提升模型的收敛速度。

2.提高模型的精度。

目前，比较常用的数据归一化的方法 Min-Max 归一化、0 均值归一化，以及其他数学函数演变而来的归一化方法，本文采取的方法是 Min-Max 归一化。

Min-Max 归一化，即对 AETA 数据作线性变换，使变换后的 AETA 数据的值大于 0，小于 1。其中，转换函数如公式(2.1)所示：

$$x^* = \frac{x - \min}{max - min} \tag{2.1}$$

其中 max 为数据的最大值，min 为数据的最小值。

本文选用样例数据进行归一化，结果如图 2.13 所示：

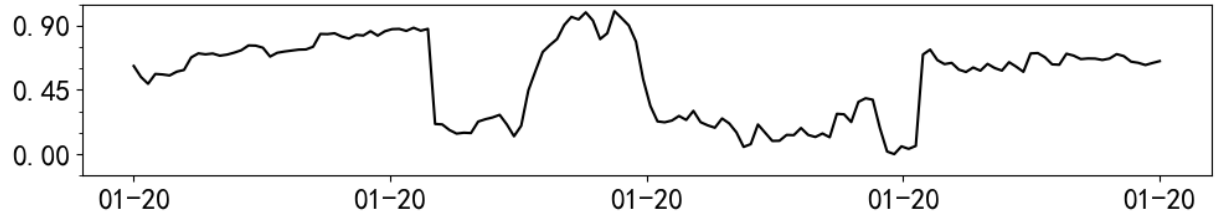


图 2.13　归一化后的波形图

2.4 本章小结

本章总共分为数据的分析和处理两部分。第一部分：首先对数据来源多分量地震监测系统 AETA、电磁探头、地声探头以及 AETA 采集数据的各分量进行了介绍。然后根据 AETA 的数据特点进行分析，讨论了由于断网导致的缺失数据处理方法和由于断电导致的缺失数据处理方法，以及补充后的完整数据的归一化处理方法。经过这些处理，可以得到后续的特征提取等工作的可用数据。第二部分，对数据的分析工作：本文基于九寨沟 7.0 级地震为例进行单台站的电磁数据和地声数据进行临震分析。

第三章　基于 AETA 数据的特征工程

在进行分类建模时，每个样本可以用 N 维的特征向量表示，N 维特征向量可以组合成不同的特征集合，而不同的特征集合携带了不同的分类信息。如果特征集合含有的分类信息可以反应样本的类别，则可以提高分类器的性能；反之，则会降低分类器的性能。因此，需要找到使得分类器在数据集中表现最优的样本特征向量的组合。

本文将会从 AETA 电磁、地声数据出发进行特征工程，如图 3.1 所示。

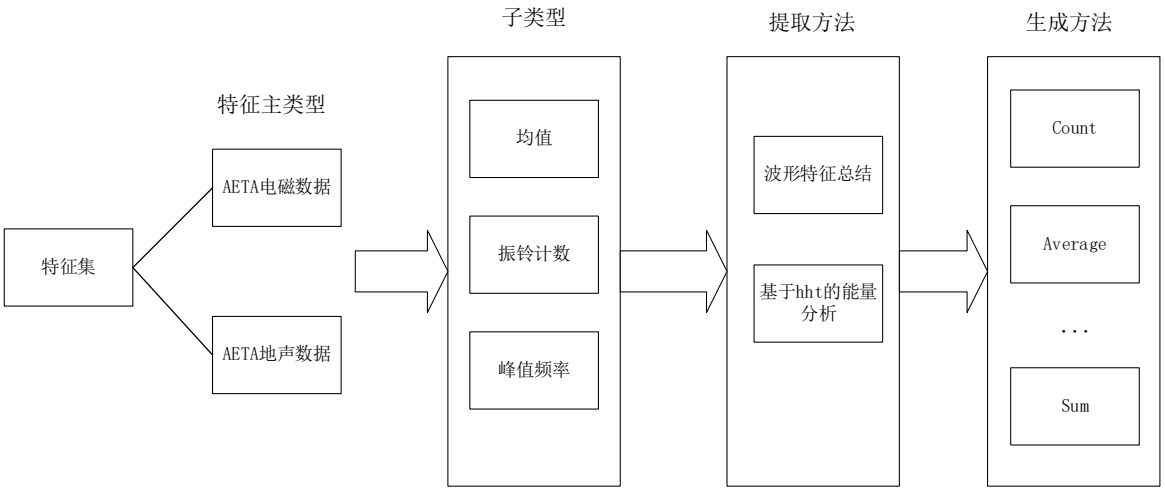


图 3.1　特征工程流程图

3.1　基于电磁数据波形的特征提取

本文根据上一节统计出的重要数据波形，利用波形相似度分析对研究区域内的地震的震前 15.10.7 天波形出现的次数进行波形统计。

AETA 台站遍布多个区域，各台站的波形也不尽相同。通过查看 AETA 的观测数据，发现多个地区多个台站 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、以及不规则波动等形态。

近似日周期的波形中，有种与日升日落几乎同步变化的日周期波形，日升时变低、日落时变高，称之为 SRSS 波，波形示意图如图 3.2 所示。

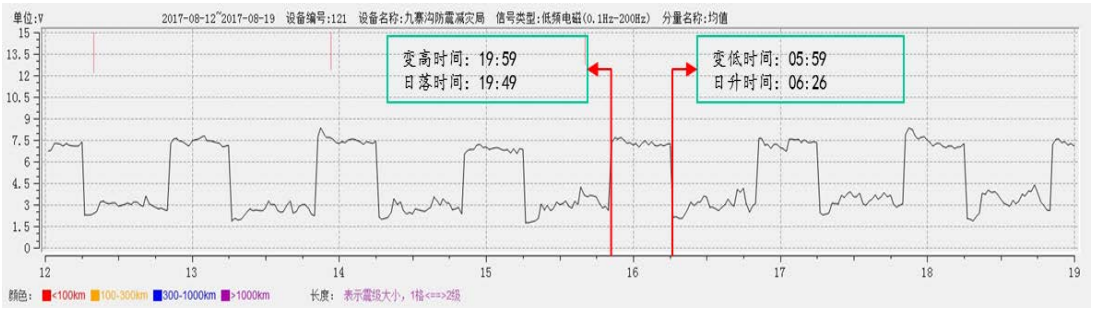


图 3.2　SRSS 波示意图

在第二章第二节已对 SRSS 波基于震例进行分析，下面本文将基于 SRSS 波以及类 SRSS 波进行波形的具体的定义与统计。

3.1.1　AETA 电磁数据的波形定义

波形 1：波形 A

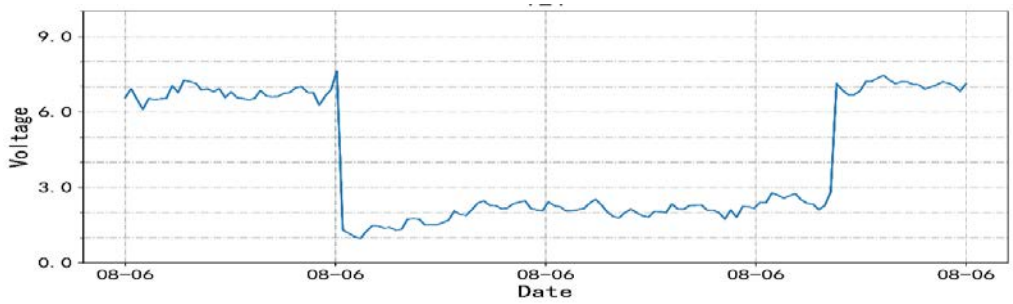


图 3.3　波形 A 示意图

从图中可以发现，震前 8 月 6 日的九寨沟防震减灾局的数据波形就是一个重要的数据形态。此波形有与日升日落几乎同步变化的日周期，日升时变低、日落时变高，本文将此定义为波形 A。

波形 2：波形 B

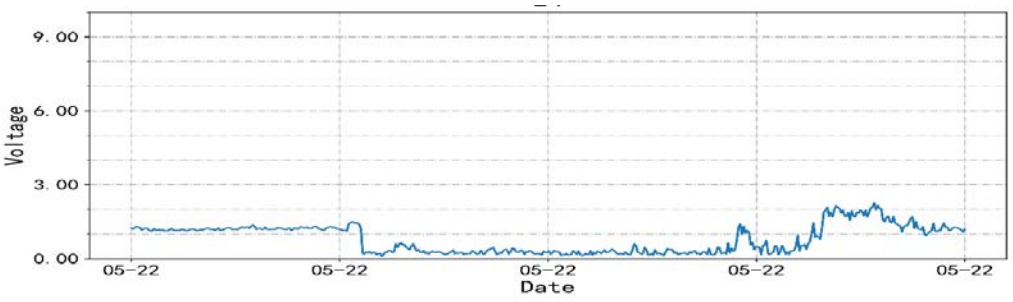


图 3.4　波形 B 示意图

从图中可以看出此波形日升时变低、日落时变高，但幅值较低，本文将此定义为波形 B。

波形 3：波形 C

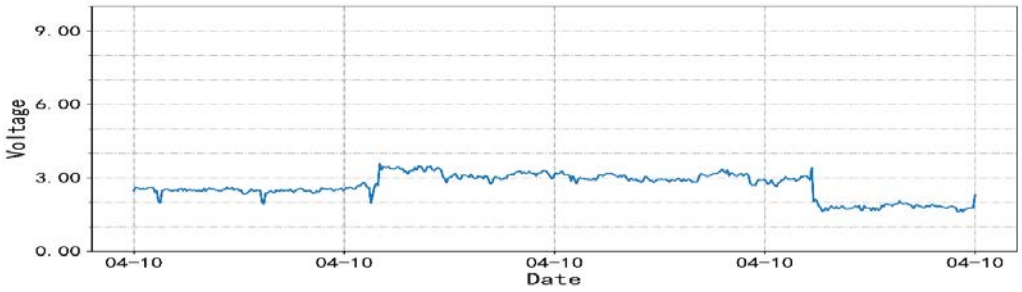


图 3.5　波形 C 示意图

从图中可以看出，该波形与 SRSS 波不同，呈现日升日落几乎同步变化的日周期，日升时变低、日落时变高的特点，本文将其定义为波形 C。

波形 4：波形 D

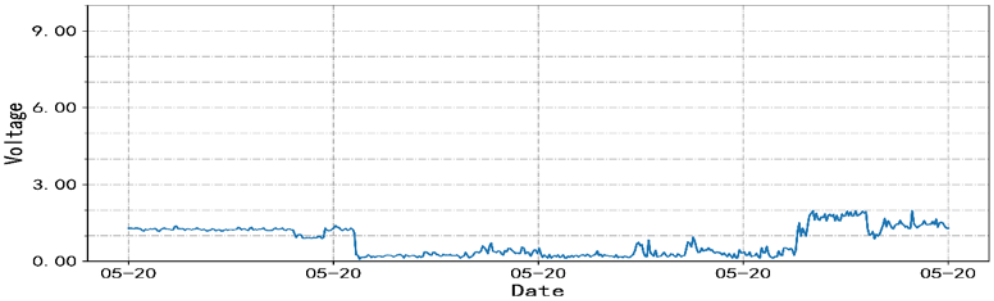


图 3.6　波形 D 示意图

从图中可以看出，该波形与 SRSS 波不同，呈现日升日落几乎同步变化的日周期，日升时变低、日落时变高，但是在中间会出现单个峰的特点，本文将其定义为波形 D。

波形 5：波形 E

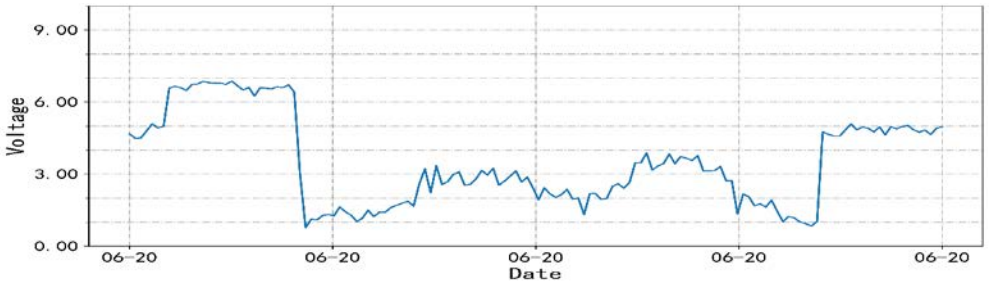


图 3.7　波形 E 示意图

从图中可以看出，该波形与 SRSS 波不同，呈现日升日落几乎同步变化的日周期，日升时变低、日落时变高，但是在中间会出现两个峰的特点，本文将其定义为波形 E。

波形 6： 波形 F

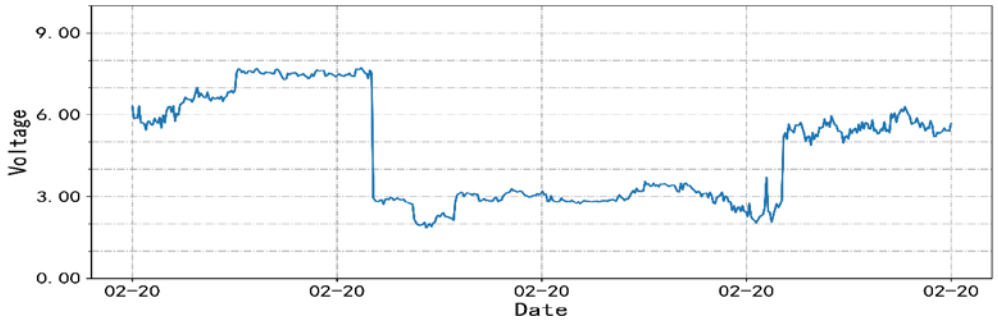


图 3.8 波形 F 示意图

从图中可以看出，该波形与 SRSS 波不同，呈现日升日落几乎同步变化的日周期，日升时变低、日落时变高,但是出现左边的幅值比右边的幅值高的特点，本文将其定义为波形 F。

波形 7：波形 G

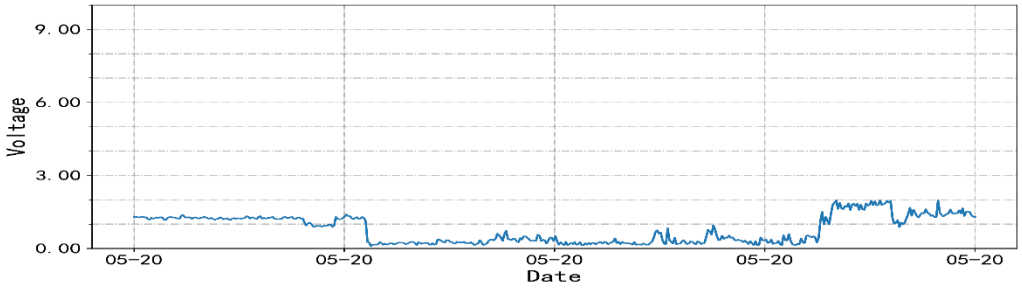


图 3.9 波形 G 示意图

从图中可以看出，该波形与 SRSS 波不同，呈现日升日落几乎同步变化的日周期，日升时变低、日落时变高,但是出现右边的幅值比左边的幅值高的特点，本文将其定义为波形 G。

3.1.2 相似度计算

1.基础波形提取

首先，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。

具体分类情况如图 3.10 所示:

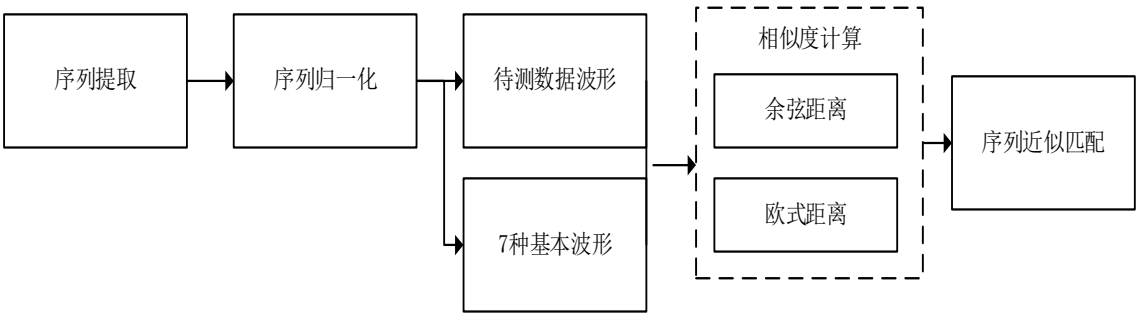


图 3.10　波形模式识别流程

2.相似度计算

提取到基础波形后，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：

$$d=\sqrt{\sum_{i=1}^{N}(x_{1i}-x_{2i})^2} \tag{3.1}$$

衡量余弦相似度的公式如下：

$$\cos\theta=\frac{\sum_{i=1}^{n}x_{1i}\times x_{2i}}{\sqrt{\sum_{i=1}^{n}(x_{1i})^2}\times\sqrt{\sum_{i=1}^{n}(x_{2i})^2}}=\frac{X_1^T\cdot X_2}{\|X_1\|\times\|X_2\|} \tag{3.2}$$

3.序列近似匹配

时间序列数据挖掘，近似波形匹配方法。

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　波形识别算法

算法：波形识别算法
输入：　基础波形 BW，目标波形　OB,.
输出：　两个时间序列之间的相似度
过程：
1: 初始化 $\tau \leftarrow \emptyset$
2: for each OBi $\in$ stations do
3:　　if η ==0 then
4:　　$\tau \leftarrow \tau \cup$ 利用余弦距离计算相似度

续表 3.1 波形识别算法

5: else if η==1 then
6: $\tau \leftarrow \tau \cup$ 利用欧式距离计算相似度
7: end if
8: end for
9: return 相似度$\leftarrow \tau$

本文以一个待测波形为例，对上述定义的波形进行相似性分析，得到的结果如图 3.11 所示：

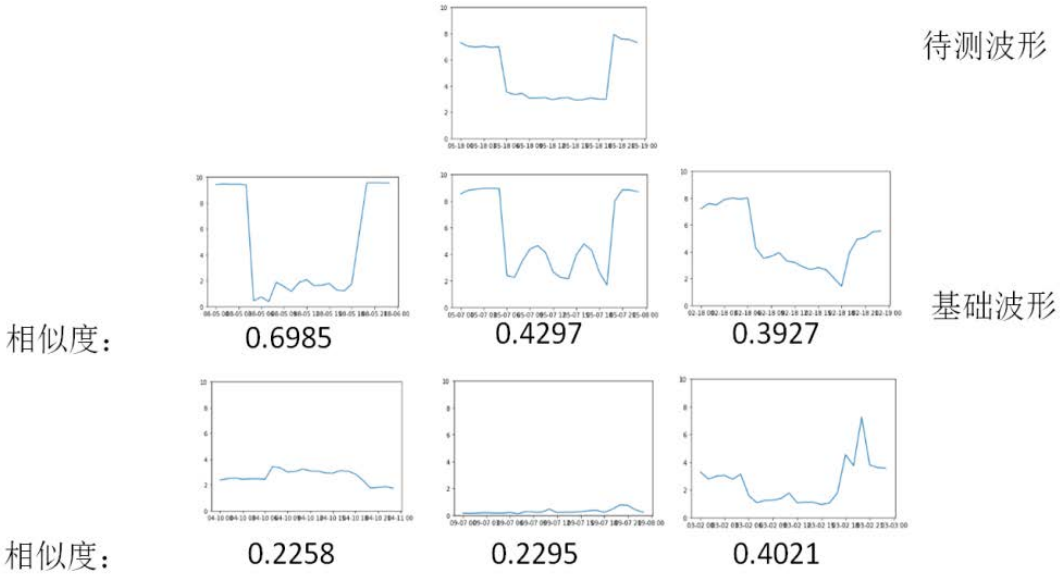


图 3.11 波形相似度分析

第一行为待测波形，第二行、第三行为已知的基础波形，下方的数字即为相似度。

3.1.3 波形特征描述

通过序列匹配结果形成对波形的描述，上述算法，可以对于任意一天的数据进行宏观上的波形描述，对震前 15 天，10 天，7 天进行特征描述，得到如下的特征表:

表 3.2 特征总结

序号	特征
1	震前 7 天电磁均值出现波形 A 的数量
2	震前 10 天电磁均值出现波形 A 的数量
3	震前 15 天电磁均值出现波形 A 的数量

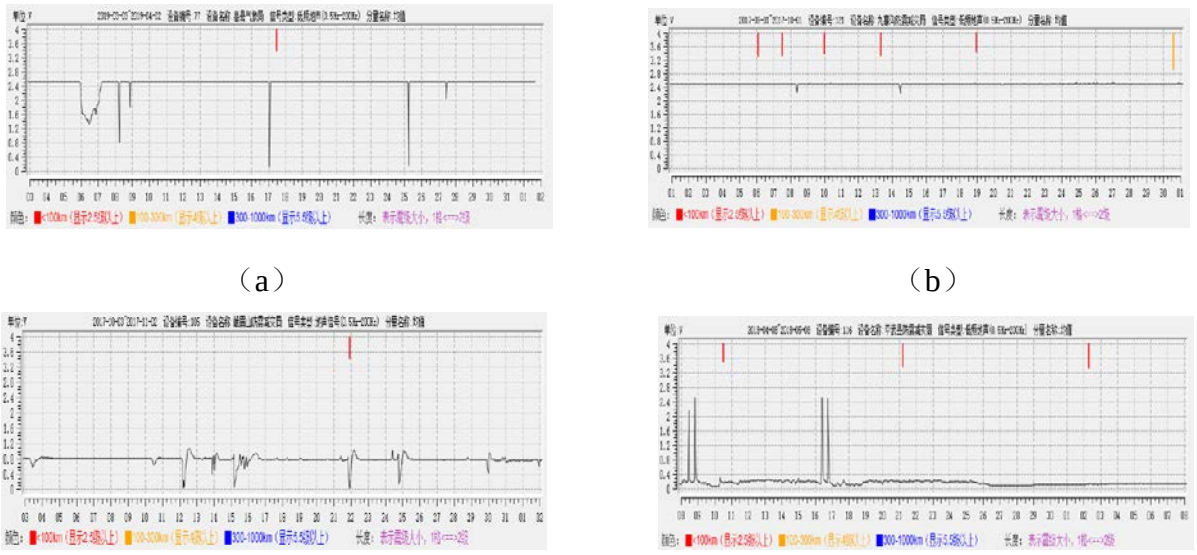
续表 3.2　特征总结

4	震前 7 天电磁均值出现波形 B 的数量
5	震前 10 天电磁均值出现波形 B 的数量
…	…
19	震前 7 天电磁均值出现非定义波形的数量
20	震前 10 天电磁均值出现非定义波形的数量
21	震前 15 天电磁均值出现非定义波形的数量

3.2　基于地声数据波形的特征提取

可以观察到地声数据的特点并非像电磁数据一样具有 SRSS 波，以及类 SRSS 波的特征；但是在地震前会现尖峰如图所示，因此，本文针对低声数据的尖峰进行特征提取。

3.2.1　地声信号数据特点



（a）　（b）

（c）　（d）

图 3.12　地声尖峰的台站波形图

图(a)表示会表示出现长的向下尖峰，表示出现短的向下尖峰图（c）表示在一天之内同时出现上下尖峰图（d）表示出现长的上尖峰。

3.2.2 地声数据特征提取算法

地声数据的尖峰提取算法如表 3.3 所示:

表 3.3 尖峰识别算法

算法： 尖峰识别算法
输入： 正常地声数据 Y，待测地声数据 X
输出： 尖峰点
1: 初始化 $\tau \leftarrow \emptyset$
2: for each Xi∈stations do
3: 计算待测点与正常点之间的距离
4: if distanace>σ
5: 则为尖峰点
6: else：
7: 为正常点
8：end if
9: end for
10: return

3.3 基于希尔伯特-黄变换的瞬时能量分析法

3.3.1 希尔伯特-黄变换方法的基本理论

希尔伯特-黄变换（简称 HHT）是 N.E Huang 等[85]提出的具有突破性意义的信号处理方法。在实际的研究中，实际采集，分析的信号很少为单分量信号，多数为多分量信号。但是绝大部分的对信号瞬时频率进行分析的方法无法用于多分量信号，只能分析单分量信号。

HHT 的突破性的意义就是能够实现传统的信号分析方法无法满足的对多分量信号进行处理分析。它可以准确的分析信号的瞬时能量在频率维度上和时间维度上的分布规律。HHT 由于对信号的滤波效果更好，以及对信号的时间和频率两个方向上的分析更加准确，可有效的用于非线性、非平稳的信号处理方面。正因为希尔伯特-黄变换在分析信号有很好的效果，现如今已经应用在各个领域，例如：非线性系统分析[86]、地震信号分析[87]、医学信号分析[88]等，在理论研究和工程应用上都有十分重要的意义。

基于 HHT 的信号分析主要分为经验模态分解(EMD)和希尔伯特变换。首先基于 EMD 方法将采集的信号进行分解，使其分解成模态函数（简称 IMF）；之后，对每个模态函数进行希尔伯特黄变换，得到每一个模态函数的 Hilbert 谱；最后，将每个模态函数的 Hilbert 谱进行汇总即生成了样本信号的 Hilbert 谱。

1.经验模态分解

由于本文采用的数据为 AETA 实时采集的电磁信号和地声信号，信号本身存在不同的波动模式，无法直接进行希尔伯特变换。对 AETA 信号进行分析时，需要对每一个波动的模式进行分析，因此在进行希尔伯特变换之前需要对样本信号进行 EMD 分解。EMD 分解具有很高的信噪比，可对 AETA 数据进行平稳化处理，以便后续进行希尔伯特变换的特征提取。

EMD 分解是信号在时间范围上进行分解，最后将信号分解成若干个分量，其中每个分量被称作内模函数(简称 IMF)。AETA 信号经过 EMD 分解后得到的分量不是都为 IMF，分解后的分量需要满足以下条件：

(1) 在 AETA 信号的持续时间内，信号与 x 轴交点的个数与信号的极值点的个数之间的差值不大于 1；

(2) 在信号持续的时间内任意时刻内，下包络线和上包络线之间的平均值为 0。

假设 AETA 信号为 X(t)，对其进行 EMD 总共三步。具体的步骤如下：

(1) 首先找到 AETA 信号存在的全部极值点，之后利用三次样条函数对之前找到的极大值点进行插值，得到上包络线，记作 $X_{max}(t)$；同理，便可得到下包络线，记作 $X_{min}(t)$。

对得到的上包络线和下包络线进行求均值，得到曲线 $m_i(t)$：

$$m_i(t) = [X_{max}(t) + X_{min}(t)]/2 \tag{3.3}$$

若 $h_1(t)$满足 IMF 所需的两个条件的，则：

$$h_1(t) = x(t) - m_1(t) \tag{3.4}$$

(3.2)式成立，所以 $h_1(t)$即 X(t)的第一个 IMF 分量。

(2) 若 $h_1(t)$不能满足 IMF 所需要的条件，那么 $h_1(t)$无法进行后续步骤。$h_1(t)$将作为原始信号，继续重复(1)，重新计算 Xmax(t)和 Xmin(t)，再利用(3.4)式计算$m_i(t)$，循环往复 k 次，直至$C_1(t) = x(t) - m_{1k}(t)$满足 IMF 所需要的条件。

(3) 从 X(t)中提取 $C_1(t)$，计算残差$r_1(t) = X(t) - C_1(t)$。

$r_1(t)$将继续作为原始信号，重复进行上述步骤，得到 n 个 IMF，直至当误差小于预定误差时，或残差无法继续提取时，结束循环。

根据以上步骤，原始信号 X(t)表示为：

$$X(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{3.5}$$

2.Hilbert 变换

对每个 IMF 做 Hilbert 变换：

$$x_i(t) = c_i(t) \tag{3.6}$$

$$Y_i(t) = \frac{1}{\pi} PV \int \frac{x(\tau)}{t-\tau} d\tau \tag{3.7}$$

其中，PV 为柯西主分量，构造解析信号$Z_i(t)$，即

$$Z_i(t) = x_i(t) + iy_i(t) = a_i^{ei\theta_i(t)} \tag{3.8}$$

其中，幅值函数$a_i(t) = \sqrt{x_i^2(t) + y_i^2(t)}$，相位函数$\theta_i(t) = \arctan(y_i(t)/x_i(t))$

通过相位函数可以得到瞬时频率为

$$w_i(t) = \frac{d\theta}{dt} \tag{3.9}$$

由式可得：$w_i(t)$表示的是时间的单值函数，即某一时刻只对应一个频率。

基于以上，则原始的给定信号可以表示为：

$$\mathrm{X(t)} = \mathrm{Re}\sum_{i=1}^{n} a_i(t)e^{i\theta_i(t)} = \mathrm{Re}\sum_{i=1}^{n} a_i(t)e^{\int w_i(t)dt} \tag{3.10}$$

式(3.9)忽略了残余项。其中，Re 表示取实部。

基于式(3.8)和式(3.9)，可得到时间、瞬时频率、瞬时幅值的三维谱图，称为 Hilbert 谱。

Hilbert 边际谱即在(0,T)时间内对$H(w,t)$进行积分：

$$\mathrm{h(w)} = \int_0^T H(w,t)dt \tag{3.11}$$

边际谱表示信号持续的时间内信号幅值的积聚，表示 AETA 信号的能量随频率的分布情况。

Hilbert 瞬时能量即对$H^2(\omega,t)d$进行积分，得到：

$$\mathrm{E(t)} = \int H^2(\omega,t)d\omega \tag{3.12}$$

Hilbert 瞬时能量表示了 AETA 信号的能量随着时间的实时变化情况。

下图为信号经过 EMD 分解的希尔伯特变换的完整程序流程图：

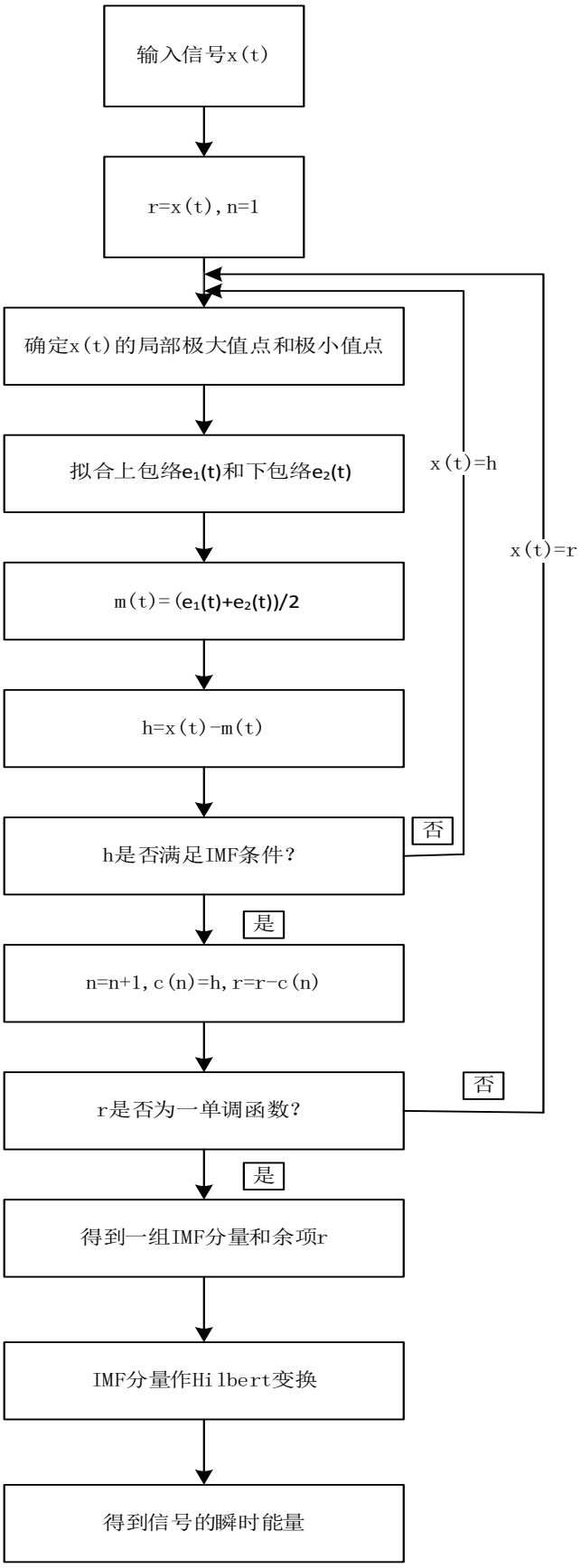


图 3.13　希尔伯特黄变换整体流程图

3.3.2 基于希尔伯特-黄变换的瞬时能量分析法

3.3.1 节分析了 HHT 变换的基本原理以及瞬时能量的定义，本节将以实际阐述 HHT 变换应用在 AETA 数据上的信号分析处理。

以 AETA 电磁数据的均值为例，对电磁数据的均值进行 EMD 分解和 Hilbert 变换，将经变换后的波形振幅进行积分，可定义 Hilbert 瞬时能量为：

$$\mathrm{h(w)} = \int_0^T H(w,t)dt \tag{3.13}$$

本文将基于 HHT 对 AETA 信号进行处理，分析 AETA 在震前的能量随时间的变化情况。

以 2017 年 8 月 7 日的九寨沟防震减灾局为例，如图 3.14 所示：

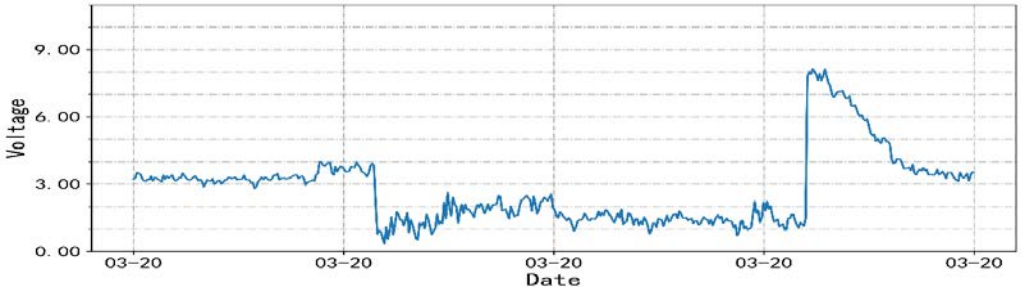


图 3.14　九寨沟防震减灾局

首先对 AETA 数据进行 EMD 分解，将信号分解成若干 IMF，结果如图 3.15 所示：

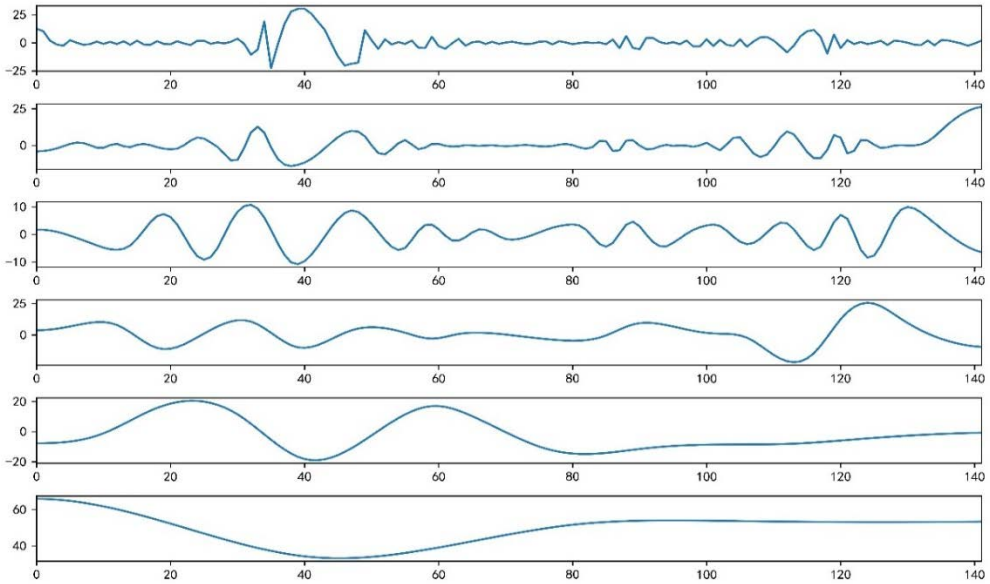


图 3.15　IMF 分量

在得到 AETA 信号所有的 IMF 分量之后，对所有的内模函数进行 Hilbert 变换。然后根据$\mathrm{h(w)} = \int_0^T H(w,t)dt$ 计算 AETA 信号每个 IMF 分量的瞬时能量。分析处理得到的瞬时能量图如图 3.16 所示。

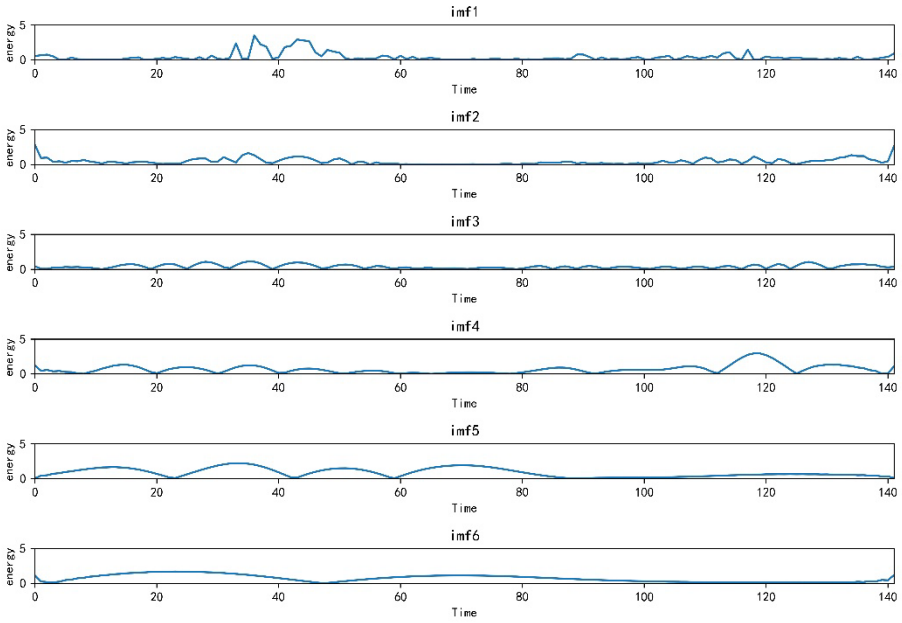


图 3.16　瞬时能量图

由图 3.16 可以得到该台站的瞬时能量随时间的分布情况，可以看出，imf1 分量的能量分析图相比于其他分量的能量分析图可以更能代表此信号的能量分析图。因此，本文选 imf1 分量的能量分析作为本文的特征，并对此进行分析。

3.3.3　基于希尔伯特-黄变换的数据频域特征提取

本节将针对 AETA 电磁、地声特征数据均值、振铃计数、峰值频率三个分量进行基于希尔伯特-黄变换的能量分析，以 2017 年 8 月 8 日九寨沟 7.0 级地震为例距震中 150km 以内的台站为例。

台站具体信息见表 3.4 所示：

表 3.4　台站信息表

序号	台站	经度/(°)	纬度/(°)	震中距/km
1	九寨沟防震减灾局（JZG）	104.25	33.26	40
2	松潘地震台(SP)	104.03	32.65	64
3	平武县防震减灾局(PW)	103.60	32.33	111
4	青川县防震减灾局(QC)	105.23	32.59	147

四个台站的地震前一天即 2017 年 8 月 7 日的波形图以及能量图如图 3.17 所示：

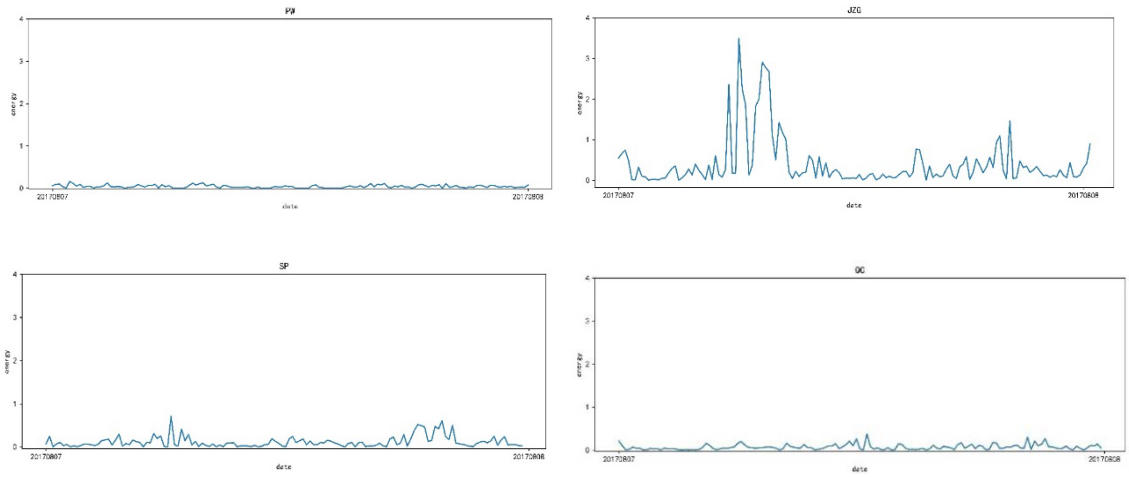


图 3.17　四个台站的震前一天能量分析图

从图中可以看出，随着台站距震源中心的距离越近，经希尔伯特-黄变化的能量谱的幅值越大，说明波形经过希尔伯特-黄变换的能量谱可作为地震前的有效特征。

本文从震前 7 天、10 天、15 天进行基于希尔伯特-黄变换的能量分析。

下面四幅图分别为九寨沟，平武，松潘，青川四个台站震前 7 天到震后 7 天的 AETA 电磁均值经过希尔伯特黄变换得到的瞬时能量图，如图 3.18 所示：

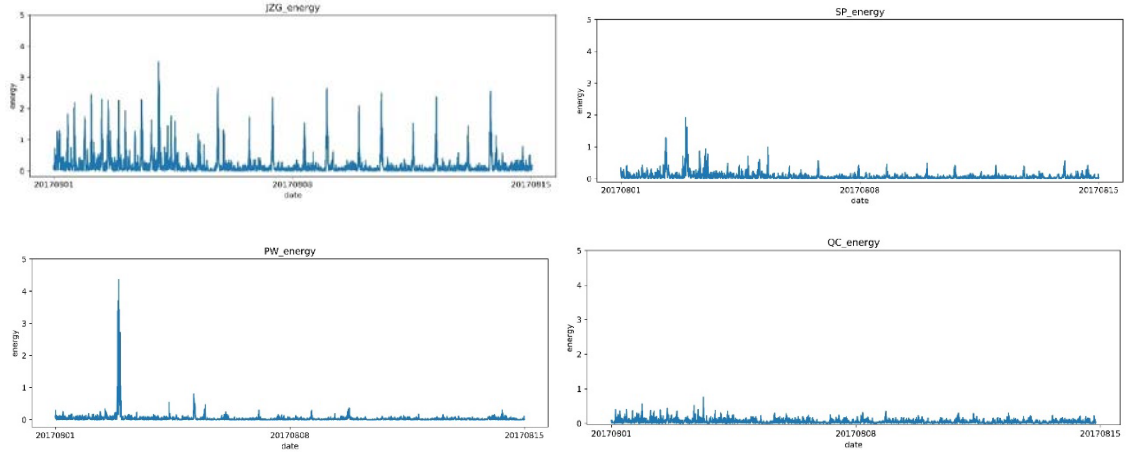


图 3.18　四个台站的震前 7 天以及震后 7 天的能量分析图

下面四幅图分别为九寨沟，平武，松潘，青川四个台站震前 10 天到震后 10 天的 AETA 电磁均值经过希尔伯特黄变换得到的瞬时能量图，如图 3.19 所示：

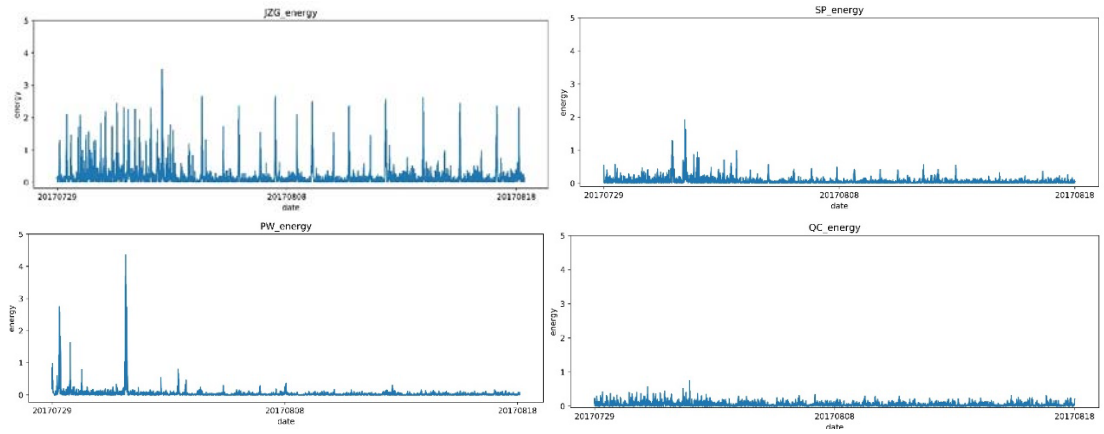


图 3.19　四个台站的震前 10 天以及震后 10 天的能量分析图

下面四幅图分别为九寨沟，平武，松潘，青川四个台站震前 15 天到震后 15 天的 AETA 电磁均值经过希尔伯特黄变换得到的瞬时能量图，如图 3.20 所示：

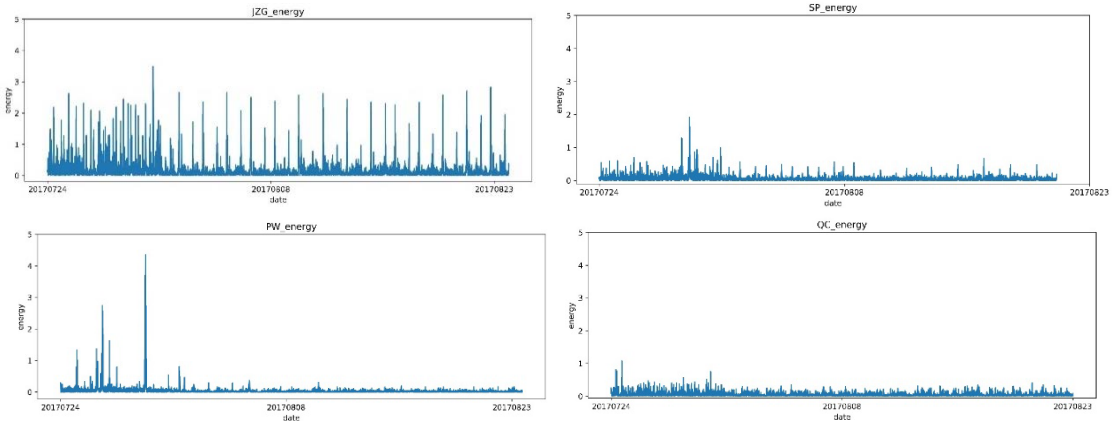


图 3.20　四个台站的震前 15 天以及震后 15 天的能量分析图

小结：

1.从图中可以看出，随着距震中的距离越远，台站的瞬时能量尖峰出现的时间越早；

2.同时发现，距离震中越近的台站，伴随“主尖峰”出现，会出现其他小的尖峰。

3.根据震前 15 天、10 天、7 天的数据，可以粗略进行判断，震前 15 天所包含的其前兆信息更加完整。

3.3.4　希尔伯特-黄变换特征总结

经过希尔伯特-黄变换得到瞬时能量，利用统计方法将特征进行量化，提取瞬时能量的最大值，均值，平均差等统计量以及瞬时能量的波峰的个数，得到的特征列表如下所示：

表 3.5　特征总结

序号	特征
1	电磁均值能量最大值
2	电磁均值能量波峰值
3	电磁均值能量波峰个数
4	电磁均值能量平均值
5	电磁均值能量标准差
6	电磁均值能量中位数
…	…
36	地声振铃计数能量中位数

3.4　本章小结

本章整体可分为两部分：第一部分，基于波形本身的特点进行分析。首先通过分析电磁数据在震前会出现几种特征波形，对震前出现的 SRSS 波及类 SRSS 波形进行编码，利用波形相似性分析，对震前 15,10,7 天的数据进行波形的特征提取。然后基于地声数据平稳，但是震前出现尖峰的特点，利用欧式距离对尖峰进行提取。第二部分，则是利用希尔伯特-黄变换对电磁数据和地声数据进行瞬时能量提取。首先利用 EMD 分解，将数据分解成 imf 分量，本文选取最能代表信号的 imf1 分量进行希尔伯特变换，将变换得到的信号瞬时能量进行了分析，得到电磁数据以及地声数据震前 15,10,7 天的瞬时能量变化。

第四章　基于 AETA 数据的特征权重评估研究

4.1　特征选择方法概述

由于地震数据存在高冗余、高噪声和高相关性的特点，使得选择合适的特征选择方法分析便成为了一项十分具有挑战性的工作。针对这一问题，相关的特征选择方法也成为研究的热点。

特征选择即在样本集的所有特征中，通过计算特征与标签之间的相关性来去除一些冗余特征的过程。一般通过特征选择后可以使机器学习模型在计算速度和精度上有所提升。一种包含四个步骤的特征选择框架[89]如下图 4.1 所示：

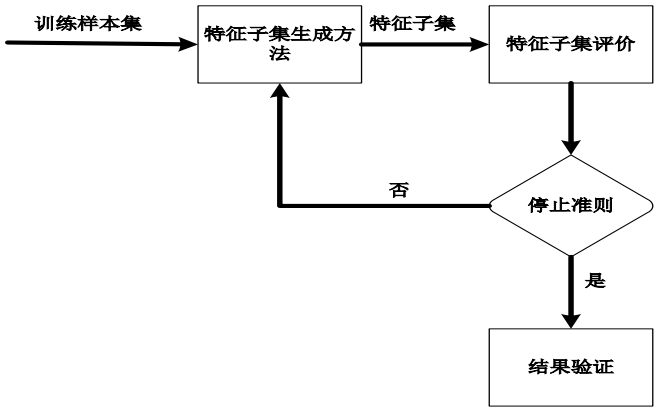


图 4.1　特征选择的过程

该框架的主要步骤为：

（1） 将特征全集按一定规则进行切分，得到多个特征子集。

（2） 通过计算特征子集的评价函数选择合理的特征子集。其中，评价函数可以是皮尔逊相关系数、互信息熵、预测准确率等表征特征有效性的指标。

（3） 遍历所有特征子集，若当前特征子集的特征有效性指标满足停止准则，则停止迭代，否则继续，若迭代完成时，仍然没有满足停止准则，则返回（1）。

（4） 对于得到的最优特征子集进行验证，具体地，把样本集的特征子集加入模型进行训练，将验证集得到的结果与特征全集在验证集上的结果进行对比。

从上述特征选择框架中可以看到，特征选择的核心有两点。第一，从特征全集中得到特征子集的过程，也即特征子集生成的策略；第二，特征子集评价的准则选取。根据这两点，可以将特征选择方法大致分为过滤式、包裹式和嵌入式这三类。

（一）过滤式特征选择算法

过滤式特征选择是指，通过一些搜索策略（如随机抽取）从特征全集中选取一部分特征作为特征子集，在该子集中计算每个特征与标签之间的评估指标（互相关系数、皮尔逊相关系数等），按评估指标高低对特征子集中的特征降序排列，最后选择评估指标最好的 TOP-k 特征进入学习算法。这种选择算法的特点是特征评估与学习算法两者之间互不相关，也不考虑特征之间的相关性，使用相互独立的判断准则，选择出最优特征子集，算法流程如下图 4.2 所示：

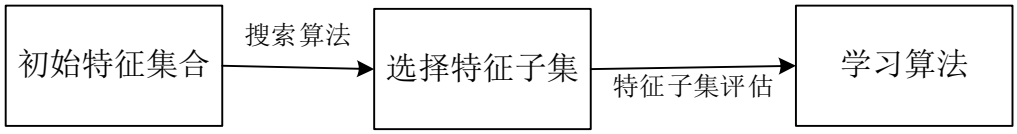


图 4.2　过滤式特征选择流程图

根据算法流程可以看出，过滤式特征选择方法的优点在于计算速度快，因为在选择过程中仅考虑特征与标签之间的关系，既不考虑特征之间的相关性也不用在选择过程中使用机器学习算法进行迭代。不过，正因如此，该方法可能会选择冗余特征进入模型，例如，在 AETA 数据中，电磁振铃计数特征和电磁频率特征与地震事件相关性都较高，但二者携带几乎相同的信息，均表征电磁信号的频率信息。使用过滤式特征选择方法有可能将二者都选入模型，从而造成冗余特征。基于此，Relief-F 算法对过滤式特征选择进行了一定改进，后文中会对该算法进行介绍。

（二）包裹式特征选择算法

包裹式特征选择算法指的是，将待选择的特征作为输入，多次训练机器学习模型（例如决策树），选择表现最好的特征集。然后将剩下的特征继续重复训练模型，直至所有待选择的特征全部都遍历过。具体算法流程如下图 4.3 所示。

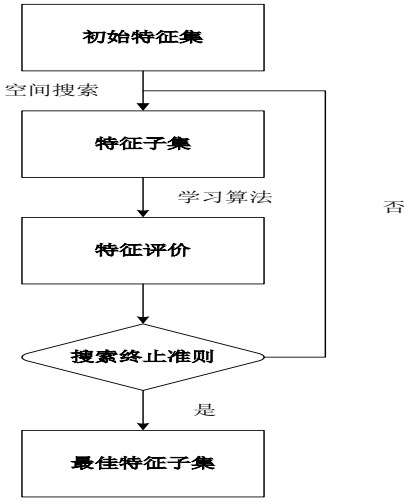


图 4.3　包裹式特征选择流程图

包裹式特征选择方法的优势在于能选择出规模较小的优化特征子集，分类精度相比于过滤式有所提高，缺点为同时提高了时间复杂度和算法复杂度，且最后选择出来的特征子集的好坏与分类模型之间存在很大的关系。

（三）嵌入式特征选择算法

嵌入式特征选择算法是指机器学习算法或者计算框架中直接包括了特征选择过程，当前的主流计算框架中均包括了嵌入式特征选择方法。例如，在 XgBoost 框架中就包括了 L1 和 L2 正则化过程，而 L1、L2 正则化本身可以通过将参数稀疏化来起到特征选择的作用。嵌入式的算法流程可总结为图 4.4：

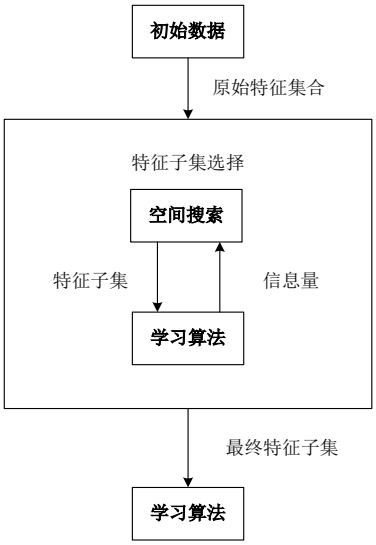


图 4.4　嵌入式特征选择算法流程图

嵌入式特征选择算法的设计结合了过滤式与封装式两者的优势，在特征选择过程中可以同时进行特征子集的搜索和评价。但嵌入式的算法复杂度也会比前两者高，并且最后选择出的特征子集好坏受到分类模型的影响。

特征选择算法的目标是寻找最优特征子集，去除不重要的特征，提高模型精度以及学习算法的性能的目标。不同特征选择的算法特征总结如表 4.1 所示：

表 4.1　特征选择算法总结

评价函数	模型特点	算法列举
过滤式	快速，与分类器独立，可扩展	Relief-F 卡方统计
包裹式	将特征选择过程与机器学习算法相结合结合依赖模型和学习算法	LVW
嵌入式	与分类器交互，较高的计算复杂能力，但分裂期依赖特征选择算法	正则化，决策树

本文采用三种方法有代表性的算法进行比较，选择出用于 AETA 数据的对地震贡献程度较大的相关特征。

4.2 用于 AETA 数据的特征选择算法的研究

4.2.1 基于 Relief-F 算法的特征选择

Relief 算法[90]是一种过滤式特征选择算法，以及基于样本学习的特征权重算法。此算法主要通过对特征在同类近邻样本与异类近邻样本中来考察其差异，并依此来度量特征的区分能力。Relief-F 算法是 Relief 算法的延伸，它具有能处理多类问题的特点以及由于采用“K 近邻”方法解决了噪声问题的特点。

该算法的具体步骤如表 4.2 所示：

表 4.2 Relief-F 算法

算法： Relief-F 算法
输入：训练样本$\mathbf{D}=\{(\boldsymbol{x_i},\boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{N}}$，xi∈RI，样本抽样次数 m，最近邻样本数为 k 输出：特征权重 W 1.初始化：置 $w_j=0$，j=1,…,I 2.for i =1 to m a) 从 D 中随机选择$\boldsymbol{x_i}$; b) $\boldsymbol{x_i}$寻找 k 个最近邻的同类样本 H_i， $\boldsymbol{x_i}$寻找 k 个最近邻的异类样本 M_i，并且 $\boldsymbol{|H_i|=\mathbf{k},\ |M_i|=\mathbf{k}}$; c) 计算每个特征的权重值： $$\boldsymbol{w_f=w_f-\frac{diff(f,\ x_i,\ H_i)}{k*m}+\frac{diff(f,\ x_i,\ M_i)}{k*m}}$$ 其中$\mathbf{diff}(\mathbf{f},\ \mathbf{x}_{\boldsymbol{i}},\ \boldsymbol{H}_{\mathbf{i}})$和$\mathbf{diff}(\mathbf{f},\ \mathbf{x}_{\boldsymbol{i}},\ \boldsymbol{M}_{\mathbf{i}})$分别表示在特征 j 上样本$\boldsymbol{x_i}$到其同类和异类近邻集合的距离和，计算公式定义如下： $$\mathbf{diff}(\mathbf{f},\ \mathbf{x}_{\boldsymbol{i}},\ \boldsymbol{H}_{\mathbf{i}})=\sum_{\mathbf{i=1}}^{\mathbf{k}}\frac{|\mathbf{x_{if}}-\mathbf{x_{if}}^{\boldsymbol{NH_j}}|}{\max\limits_{\mathbf{i=1,2\ldots N}}\{\mathbf{x_{if}}\}-\min\limits_{\mathbf{i=1,2\ldots N}}\{\mathbf{x_{if}}\}}$$ $$\mathbf{diff}(\mathbf{f},\ \mathbf{x}_{\boldsymbol{i}},\ \boldsymbol{M}_{\mathbf{i}})=\sum_{\mathbf{i=1}}^{\mathbf{k}}\frac{|\mathbf{x_{if}}-\mathbf{x_{if}}^{\boldsymbol{NM_j}}|}{\max\limits_{\mathbf{i=1,2\ldots N}}\{\mathbf{x_{if}}\}-\min\limits_{\mathbf{i=1,2\ldots N}}\{\mathbf{x_{if}}\}}$$ 其中$\mathbf{x_{if}}^{\boldsymbol{NH_j}}$表示样本的第 j 个同类近邻的第 f 个特征值及$\mathbf{x_{if}}^{\boldsymbol{NM_j}}$样本$\mathbf{x}_{\boldsymbol{i}}$的第 j 个异类近邻的第 f 个特征值。

4.2.2　基于 LVW 算法的特征选择

LVW（Les Vegas Wrapper）是一个典型的包裹式特征选择方法。LVW 是在拉斯维加斯方法（Las vegas method）的基础上采用随机策略来进行子集搜索，并根据最终分类器的误差作为特征子集的评价准则。

算法具体的步骤如表 4.3 所示：

表 4.3　LVW 算法

算法：LVW 算法
输入：数据集 D；特征集 A；学习算法 ε；停止条件控制参数 T； 输出：特征子集 A* 过程： 1.E=∞； 2.d=|A|； 3. A*=A; 4.t=0; 5.while t<T 6.　随机产生特征子集 A’ 7.　d’ = |A’| 8.　E’ = CrossValidation(ε(D A’)) 9.　if (E’≤E) and (d’<d) 10.　　t=0 11.　　E=E’ 12.　　d=d’ 13.　　A* =A’ 14.　else 15.　　t=t+1 16.　end if 17.end while

其中 E’ = CrossValidation(ε(D $^{A'}$))是通过在数据集 D 上，使用交叉验证法来估计学习器 ε 的误差，此误差是在只考虑特征子集 A’的情况下得到的，即是特征子集 A’上的误差，若 A’上的误差比当前 A 上的误差更小，或误差很小但 A’中包含的数更少，则把特征子集 A’保留下来。

4.2.3 基于随机森林的特征选择

随机森林是一种集成学习方法。集成学习一般可以归纳为串行集成方法和并行集成方法。串行集成方法，又被称为 Boosting，是加法原理叠加若干基学习器，从而集成得到一个强学习器，典型的串行集成方法有 GBDT、AdaBoost 等。并行集成方法，又被称为 Bagging，是通过若干基学习器的投票原理来集成得到一个强学习器，随机森林就是典型的并行集成学习方法。相比较于 Boosting 算法，随机森林对于噪声数据和存在特征值的数据有更好的鲁棒性，在一些数据集上，泛化能力更强，并且具有比较快的学习速度，其变量的重要性度量可以用来对高维数据进行特征选择，现如今已经被广泛用在分类，特征选择等问题中。

随机森林由多个决策树{h（x，Φi），I =1,2,…,k}组成，整个算法主要包括决策树生长和结果投票两个部分。在决策树生长过程中，为了保证集成学习器的泛化能力以及，每一棵树都使用不完全相同的数据进行训练。具体地，应用 bootstrap 重采样，从大小为 N 的数据集中随机有放回地抽取 K 个样本，利用每次抽取的 K 个样本生成一个决策树，重复 k 次，生成 k 个不同的决策树。在结果投票部分，利用 k 个决策树对样本预测结果的投票结果作为强分类器预测结果。

在 bootstrap 抽取样本时，由于是有放回的随机抽样，所以从 N 个样本中抽取 K 个，重复 k 次后，仍然会有约三分之一的样本从来没有被抽到过，这是因为，样本在 k 次采样中都没有被抽到的概率可表示为$(1-k)^k$，当 k 足够大时有：

$$\lim_{\mathrm{k}\to\infty}\left(1-\frac{1}{\mathrm{k}}\right)^{\mathrm{k}}=\frac{1}{e}\approx 0.368 \tag{4.1}$$

这部分从未被抽取到的样本常被称作袋外样本(OOB)。

在进行特征选择，需要一个指标进行评价特征分类的好坏。一般的评价函数有信息、基尼系数、卡方检验。本文采取基尼系数对特征的分类进行评价。设 AETA 信号总共有 K 类特征，其中第 i 类的概率为p_i，则得到的基尼系数为

$$\mathrm{Gini(p)}=\sum_{i=1}^{K}p_i(1-p_i)=1-\sum_{i=1}^{K}p_i^2 \tag{4.2}$$

从式中可以看出，Gini 系数与样本的概率之间成反比。因此，若 Gini 系数越大，说明p_i越小，分类的结果越不好，无法识别表现优异的特征；反之，Gini 系数越小，说明p_i越大，分类的结果越好，表现优异的特征更能被选择出来。

具体算法实现如下图 4.5 所示：

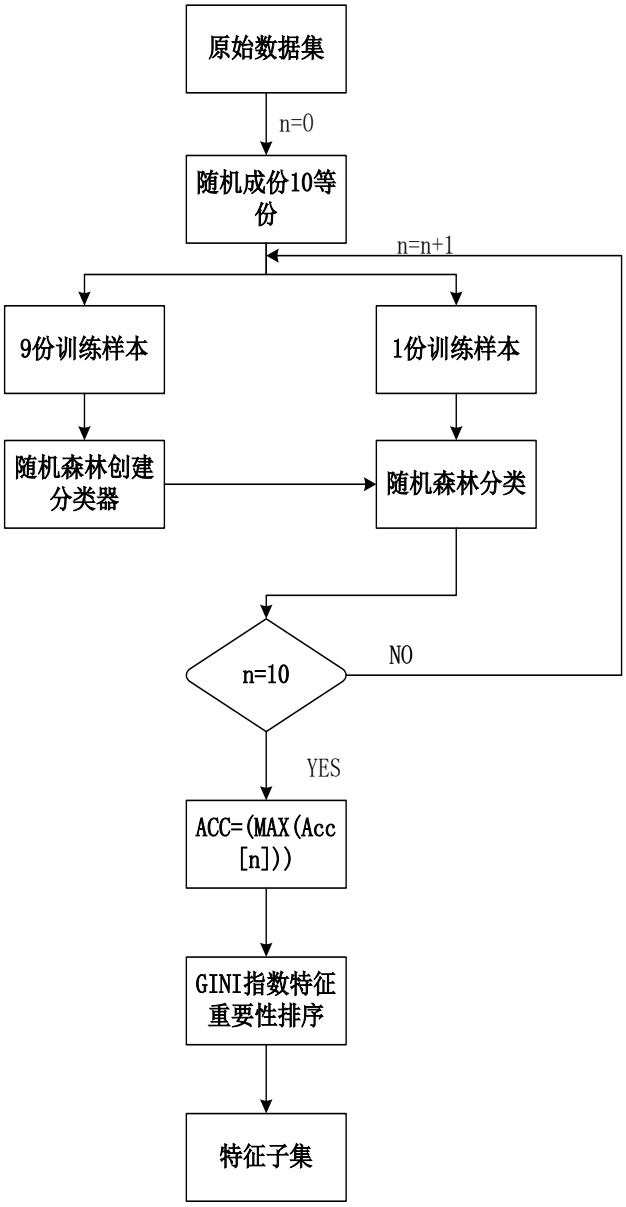


图 4.5　随机森林特征选择流程图

4.2.4　基于 AETA 数据的特征选择算法对比

ReliefF 算法、LVW 算法和随机森林分别是过滤式、包裹式、嵌入式特征选择的代表算法，其各有优缺点。但基于 AETA 数据的数据量大，会出现由于不可抗拒的因素而导致数据缺失等特点，本文将此三种特征选择算法进行了各方面的性能对比，具体的性能对比如表 4.4 所示：

表 4.4　三种特征选择算法性能对比

	Relief-F	LVW	随机森林
数据量大	×	×	√
数据缺失敏感度低	×	×	√
泛化能力	×	×	√
分类性能	√	√	√
时间复杂度	√	×	×

基于 AETA 数据的特点，本文总共提取了几十种特征，但其中含有一些冗余特征需要去除，否则会导致由于特征数量太多导致设计分类器计算开销太大，而使分类性能较差。如表 4.4 所示，随机森林算法可同时满足 AETA 数据的高维数据，数据存在缺失等特点以及分类精度不会减少特别多、选择后的结果不影响类分布，不会影响后续模型训练的结果。因此，本文最终选择基于随机森林对基于 AETA 数据所提的特征进行特征选择。

4.3　基于 AETA 数据和随机森林算法的特征权重评估

4.3.1　实验过程

1.特征集合情况概况

本文采用第三章所介绍的三种方法建立特征集合。方法一是根据提取统计震前 15.10.7 天的波形出现的次数建立特征集合 A。方法二是基于地声信号特点提取的在震前 15.10.7 天提取的尖峰数量建立特征集合 B。方法三是根据希尔伯特-黄变换提取的瞬时能量建立特征集合 C。具体特征如表 4.5 所示：

表 4.5　特征列表

对比项	特征集合 A	特征集合 B	特征集合 C
具体特征	震前 15 天电磁均值波形 A 出现的次数…	震前 15 天地声均值出现波峰的次数…	震前 15 天电磁均值的最大瞬时能量…
特征个数	21	9	36

2.随机森林模型参数

在 4.2 节通过对 relief-F 算法、LVW 算法和随机森林算法三种方法进行了研究对比，发现随机森林更适用于 AETA 数据所提取的特征，因此采用随机森林算法建立对本文所提取特征进行特征权重评估的模型，其中模型的关键参数如表 4.6 所示：

表 4.6　模型关键参数

参数	含义	详情
n_estimators	树的数量	40
criterion	树节点分裂的依据	Gini 系数
max_depth	设置树的最大深度	4
max_features	单个决策树使用特征的最大数量	所有特征
min_samples_leaf	叶子节点所含最小样本数	3
n_jobs	使用处理器的数量	-1（没有限制）
bootstrap	树生长是否有放回采样	是
其他参数		默认

4.3.2　建模结果与分析

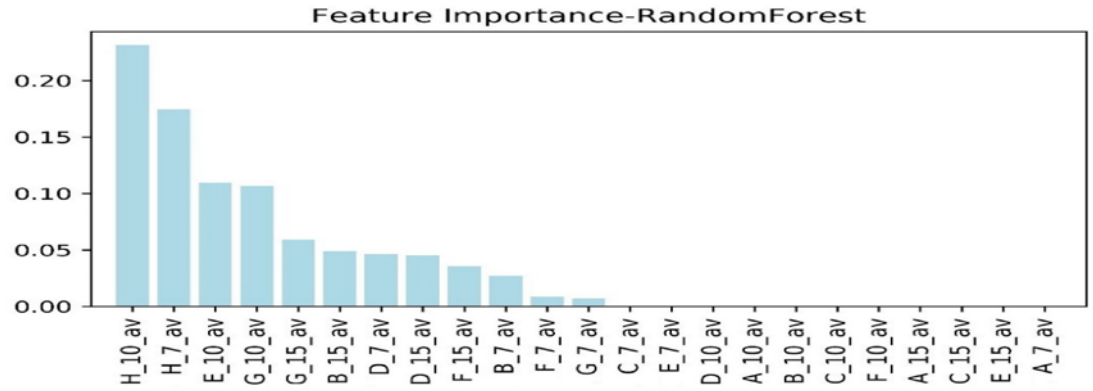
在 4.3.1 小节中所述的实验参数下，分别将区域 1 和区域 2 的数据集以 5:1 的比例切分为训练集和验证集。区域 1 中验证集有 14 个地震事件，区域 2 中验证集有 15 个地震。据此，得到 2 个区域的 3 组特征集合在模型中的表现如表 4.7 所示，可以看到，基于希尔伯特黄变换得到的特征集 C 的预测效果略优于特征集 A 和特征集 B。

表 4.7　三组特征集合在测试集上的结果对比

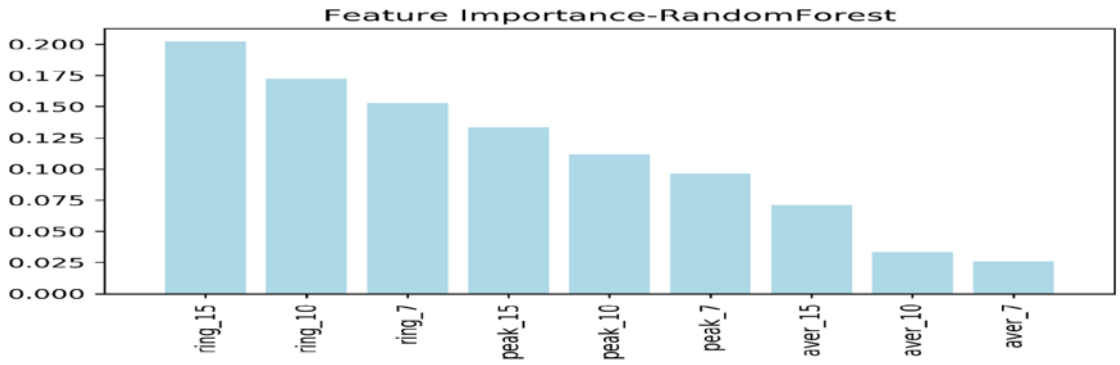
	指标	特征集合 A	特征集合 B	特征集合 C
区域 1	Precision	0.57	0.67	0.73
	Recall	0.34	0.23	0.42
	F1-score	0.43	0.34	0.53
区域 2	Precision	0.65	0.72	0.72
	Recall	0.31	0.23	0.43
	F1-score	0.42	0.34	0.54

图 4.6 和图 4.7 分别给出了三特征集合在区域 1 和区域 2 中的重要性排序，表 4.8 和表 4.9 分别表示区域 1 和区域 2 筛选后的特征列表。在决策树类的模型训练过程中，

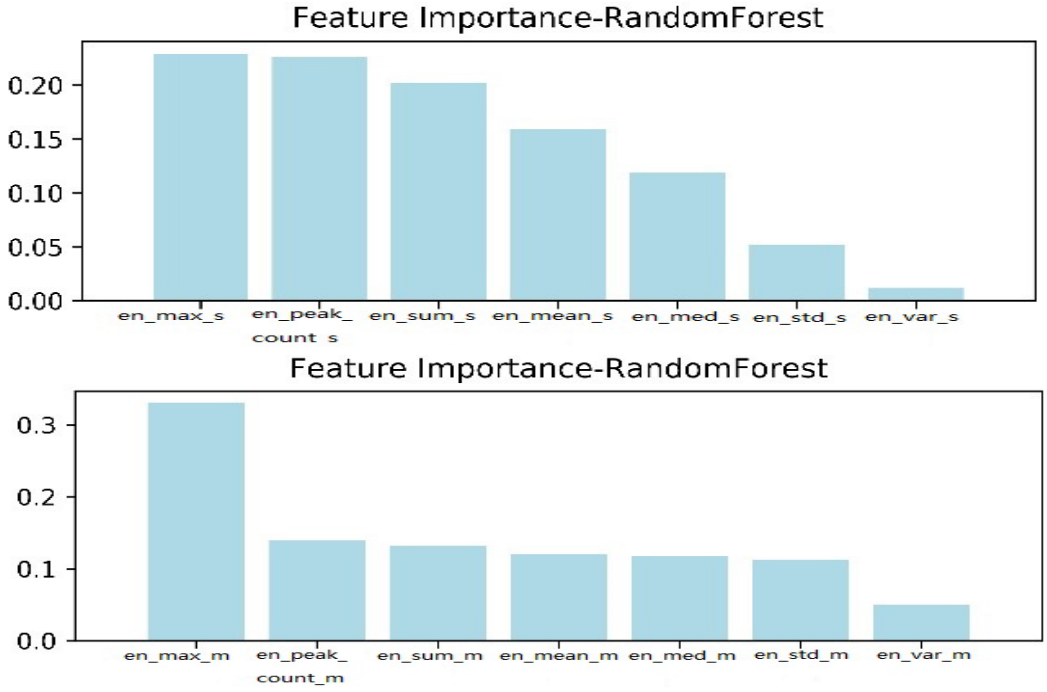
根据某个特征分裂后，数据集的基尼不纯度会降低，特征重要性是指，该特征在决策树分裂过程中导致的基尼不纯度下降总和。换言之，若在训练过程中，特征被选中次数越多，则该特征重要性越高，若某个特征为被选中过，则该特征的重要性为 0。



(a)



(b)



(c)

图 4.6　区域 1 的特征重要性排序

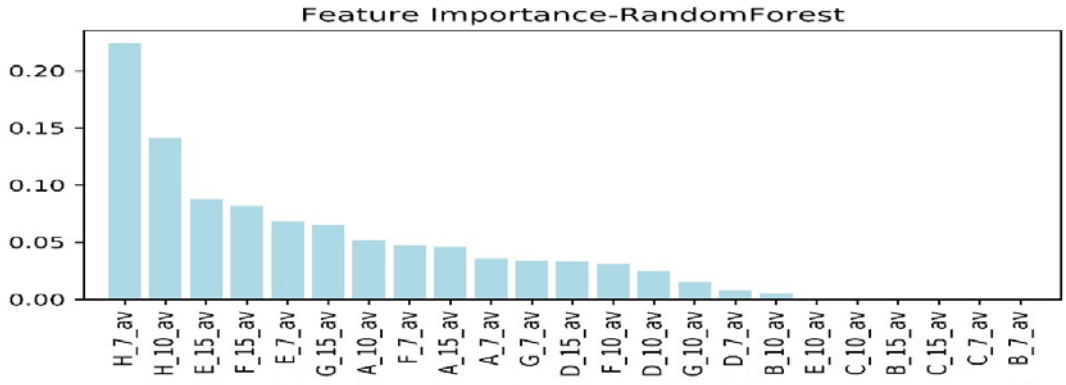
其中图 a 为特征集合 A 中的部分特征，图 b 为特征集合 B 的特征，图 c 为特征集合 c 中的特征。

可以看到，在区域 1 中，特征集合 A 和特征集合 B 的最大特征重要性都在 0.2 左右，而特征集合 C 的最大特征重要性为 0.35 左右，这也能在一定程度上解释表 4.7 的实验结果中特征集合 C 的表现最优的原因。

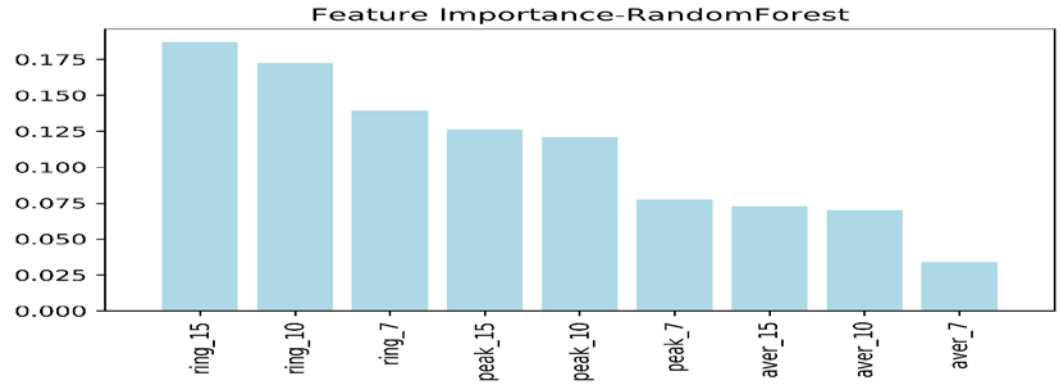
表 4.8　区域一筛选后的特征列表

特征	类型	含义
en_ max_m	连续	电磁瞬时能量最大值
en_ max_s	连续	地声瞬时能量最大值
en_peak_s	连续	地声瞬时能量峰的数量
ring_15	离散	震前 15 天振铃计数尖峰数量
ring_10	离散	震前 10 天振铃计数尖峰数量
en_std_m	连续	电磁瞬时能量标准差
en_var_m	连续	电磁瞬时能量方差
peak _15	离散	震前 15 天峰值频率尖峰数量
peak _10	离散	震前 10 天峰值频率尖峰数量
W_H _10	离散	震前 10 天其他波形的数量
en_sum_s	连续	地声瞬时能量和
en_mean_s	连续	地声瞬时能量平均值
en_median_s	连续	地声瞬时能量中位数
en_std_s	连续	地声瞬时能量标准差
en_var_s	连续	地声瞬时能量方差
W_A_15	离散	震前 10 天波形 A 的数量
W_A_7	离散	震前 7 天波形 A 的数量
W_B_10_	离散	震前 10 天波形 A 的数量
W_C_7_	离散	震前 7 天波形 C 的数量
W_F_10	离散	震前 10 天波形 F 的数量

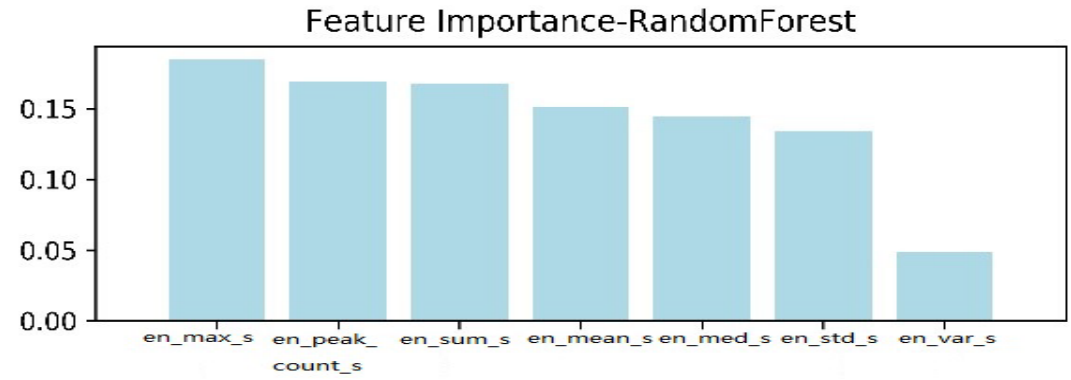
区域二的特征权重评估如图 4.7 所示：

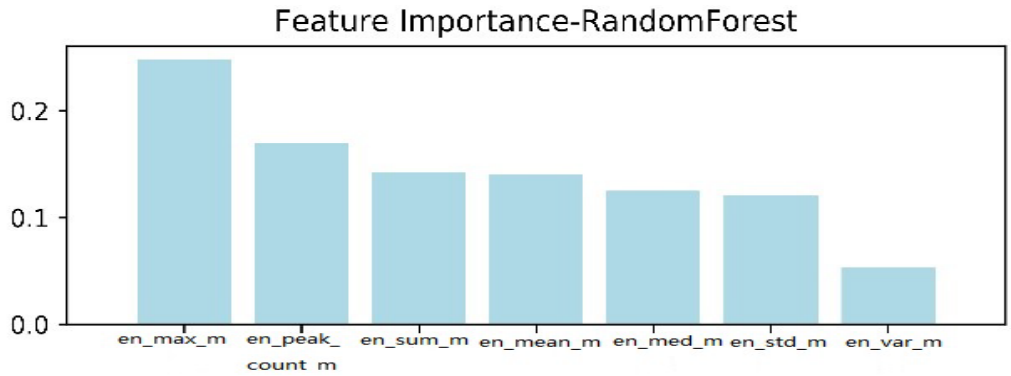


(a)



(b)





(c)

图 4.7　区域 2 的特征重要性排序

其中图 a 为特征集合 A 中的部分特征，图 b 为特征集合 B 的特征，图 c 为特征集合 c 中的特征。

对于区域 2 的特征重要性排序结果，可以发现，三个特征集合中最大特征重要性均在 0.25 左右，但是特征集合 A 中有大量特征重要性为 0 的特征，这说明在该特征集合中存在较多冗余特征，对预测结果没有贡献度；而特征集合 C 中的所有特征重要性均大于 0.1，说明在该特征集合中所有特征都对预测结果有一定贡献度。这也说明了表 4.7 中的实验结果的合理性。根据特征重要性排序，本文在三个特征集合中综合选取了平均排名最靠前的 10 个特征作为入模特征。

表 4.9 区域二选择后的特征列表

特征	类型	含义
en_ max_m	连续	电磁瞬时能量最大值
W_H_7	离散	震前 7 天的其他波形数量
en_sum_m	连续	电磁瞬时能量和
en_mean_m	连续	电磁瞬时能量平均值
en_median_m	连续	电磁瞬时能量中位数
en_std_m	连续	电磁瞬时能量标准差
en_var_m	连续	电磁瞬时能量方差
p _15	离散	震前 15 天峰值频率尖峰数量
p_10	离散	震前 10 天峰值频率尖峰数量
W_H _10	离散	震前 10 天其他波形的数量
en_sum_s	连续	地声瞬时能量和
en_mean_s	连续	地声瞬时能量平均值
en_median_s	连续	地声瞬时能量中位数
en_std_s	连续	地声瞬时能量标准差
en_var_s	连续	地声瞬时能量方差
W_A_15	离散	震前 10 天波形 A 的数量
W_A_7	离散	震前 7 天波形 A 的数量
W_B_10_	离散	震前 10 天波形 A 的数量
W_C_7_	离散	震前 7 天波形 C 的数量
W_F_10	离散	震前 10 天波形 F 的数量

4.4 本章小结

本章首先分析了进行特征选择的原因以及一般流程，并根据不同的评价函数分类的三种特征选择算法：过滤式、包裹式、嵌入式特征选择算法的原理及优缺点进行了分析比较。然后对三类特征选择的代表算法 relief-F 算法，LVW 算法，以及随机森林并对比其优缺点，选择出最适用于 AETA 数据的随机森林进行特征权重评估。最后，基于第三章提取的特征利用随机森林对所提取的包含 66 种特征的特征集进行了权重评估，筛选出了对地震预测贡献度最大 en_peak_max、ring_15、en_sum 等 20 个特征。

第五章　基于 AETA 数据的地震预测模型研究

5.1　基于 AETA 数据的地震预测框架

迄今为止，国内外学者对地震的发生机理尚有不同言论，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。

本章将在上述章节的基础上进行地震预测模型的研究，从 AETA 电磁数据和地声数据的特点出发，对电磁数据的波形形状的提取，地声波峰的提取，以及基于希尔伯特-黄变换的能量分析进行特征的提取。在进行特征空间构造后，由于出现冗余的特征会导致模型的计算开销太大，导致模型的性能变差。因此利用随机森林对冗余以及非有效的特征集合进行筛选，选择出对地震贡献度更大的特征。基于以上特征作为模型的输入，本章将对地震震级以及地点的预测模型进行研究，以及验证。图 5.1 即为预测模型流程图：

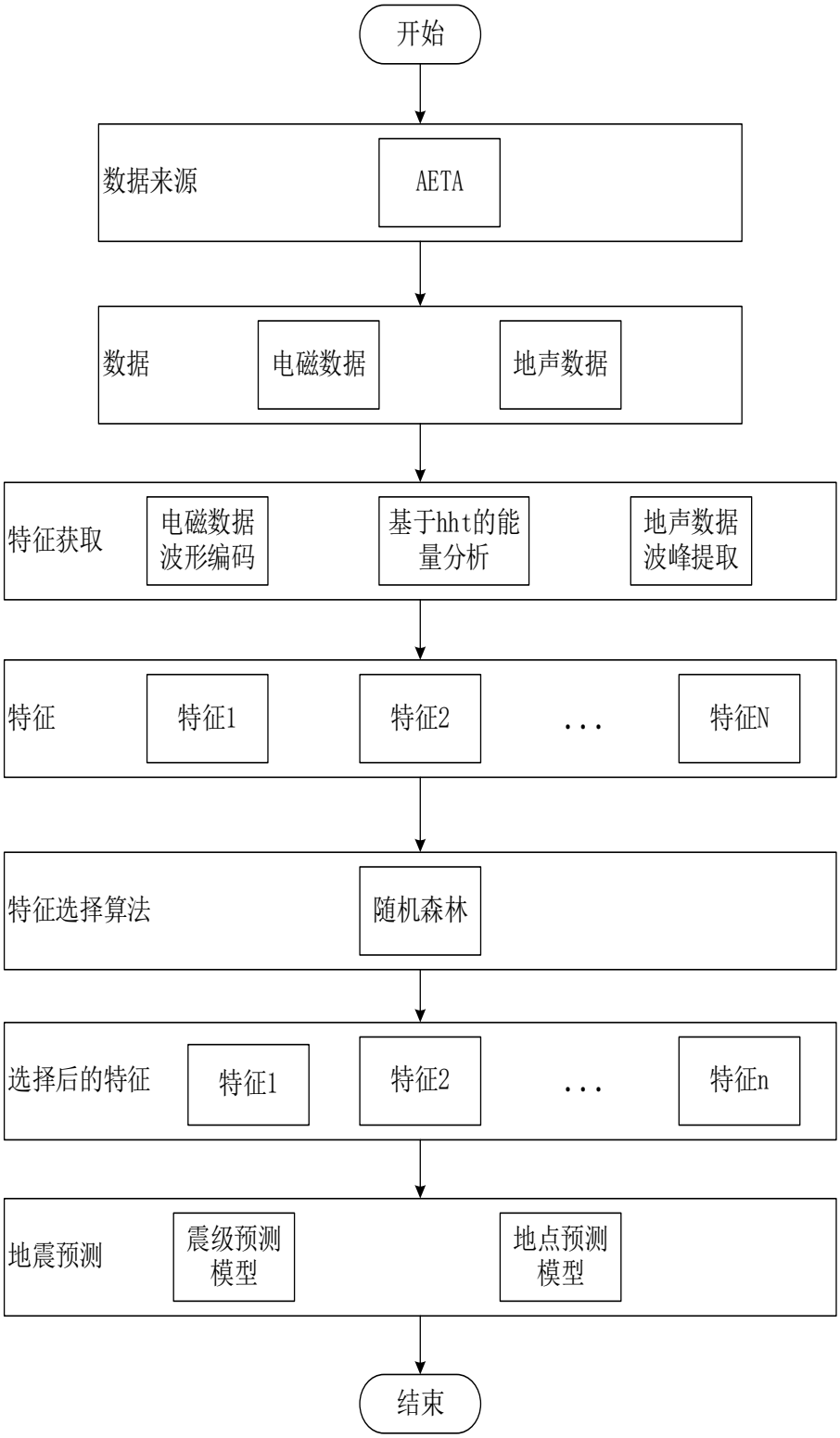

开始
数据来源
AETA
数据
电磁数据
地声数据
特征获取
电磁数据
波形编码
基于hht的能
量分析
地声数据
波峰提取
特征
特征1
特征2
...
特征N
特征选择算法
随机森林
选择后的特征
特征1
特征2
...
特征n
地震预测
震级预测
模型
地点预测
模型
结束

图 5.1　预测模型流程图

5.2　基于 AETA 数据的震级预测模型研究

5.2.1　CART 算法

CART 算法的提出是机器学习、数据挖掘领域的一个重要里程碑。CART 模型的建立可以理解为二元递归规划，最后为优化的二叉树。CART 模型建立，需要进行以下两个步骤：

（1）CART 决策树的构造。首先，从根节点利用自上到下的递归，依次在每一个节点选择分支属性。然后，在选择的分支属性之中再一次进行选择最佳分支。在这之中，所采用的方法，一般为双化指数、有序双化指数、基尼指数、LS 或 LAD 等作为度量“纯净度”的依据。

（2）CART 决策树的剪枝。在第一步的构造树的过程中，可能会发生过拟合的现象，所以需要进行下一步—树的剪枝。剪枝通常情况下包括：前剪枝，后剪枝。前剪枝为在第一步树的创建中，便可以知道需要修剪哪些节点，这些节点便可以停止生长。后剪枝为在形成一个完整的树之后，再进行修剪分支。

两者进行比较，前剪枝的优点是节约树的构造过程中的开销，后剪枝的优点是产生更优的结果。剪枝的方法一般有：最小误差剪枝、代价复杂性剪枝等。

CART 分类算法具体计算过程如图 5.2 所示：

若需要得到更好的结果，则需要对 CART 决策树进行剪枝。具体算法如表 5.1 所示：

表 5.1　CART 决策树剪枝算法

算法 5.1：　CART 剪枝算法
输入：未剪枝的决策树 T_0
输出：剪枝后的决策树 Tα
过程：
1.初始化：k=0,T= T0,α=+∞
2.自顶向下计算 $C(T_t)$,
3.对 g(t)=α 的内部结点 t 若多数表决通过则留此类，否则剪枝，最终得到树 T.
4.设 k=k+1,αk=α,Tk=T.
5.如果 T_k 是由根节点及两个叶节点构成的树，令 $T_k=T_n$ 否则，将重新回到步骤 2.
6.对子序列 $T_0,T_1,\ldots,T_n$ 进行交叉验证，选择最优子树 Tα.

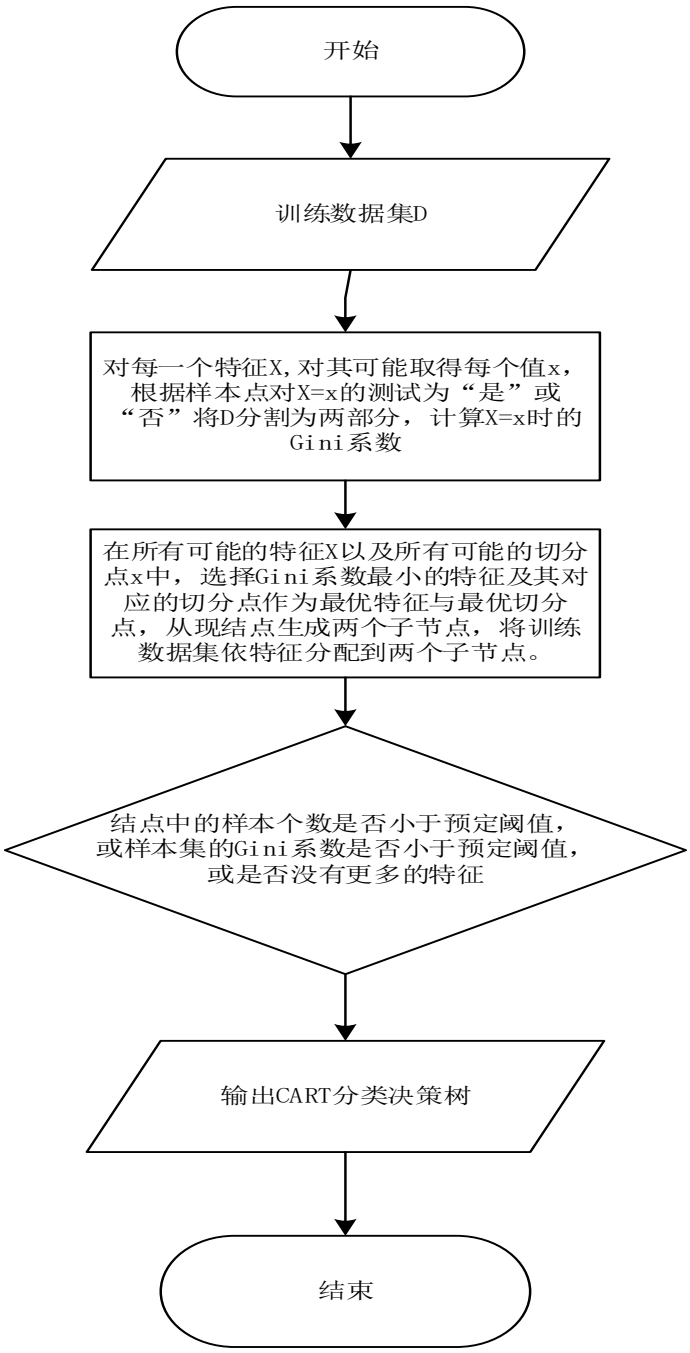


图 5.2　CART 分类决策树生成算法流程图

5.2.2　支持向量机

支持向量机（简称 SVM）已成为热点的分类技术。此技术具有坚实的统计学理论基础，同时在计算机学习、模式识别，预测预报等许多的实际应用中展示了实际的效用。SVM 可避免维灾难现象，因此用在高维数据中。

1 支持向量机原理

SVM 是一种在特征空间中间隔最大的分类器，由于包括核函数，使其不仅为线性分类器，同时也可作为非线性分类器。

设给定训练样本 T={(x1,y1),(x2,y2),...,(xN,yN)}，其中$\mathbf{x}_i \in X = R^n$，$y_i \in Y = \{-1,1\}$,i=1,2,…,N。如果数据集中所有数量均可被某超平面进行正确的分类，且边缘最大化，即与最近的异类向量之间相距最远的平面，则称作最优超平面。其中，支持向量可定义为：在所有异类向量中，距离超平面最近的即为支持向量。

SVM 是基于线性可分的最优分类面，记作$(w \cdot x) + b = 0$，将超平面进行归一化，使数据满足(5.1)式：

$$y_i(w \cdot x + b) \geq 1, i = 1, \dots, l \tag{5.1}$$

由于超平面和支持向量的距离为$\frac{1}{||w||}$，两个支持向量之间的距离为$\frac{2}{||w||}$，此问题等价于最小化$\frac{1}{2}||w||^2$.在考虑允许拟合误差的情况下，引入松弛因子$\varphi_i \geq 0$，$i = 1, \dots, l$，

超平面的约束变为：

$$y_i(w \cdot x + b) \geq 1 - \varphi_i, i = 1, \dots, l \tag{5.2}$$

优化目标转换为最小化

$$\emptyset(w, \varphi_i) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\varphi_i \tag{5.3}$$

上式的$\frac{1}{2}||w||^2$使特征被分为的类别差距变大，$C\sum_{i=1}^{l}\varphi_i$可让误变小。其中，常数 C>0，用以控制对 ε 的惩罚程度。

为了优化目标，将上式加入拉格朗日函数，使其变成：

$$L(w, b, \varphi_i, \alpha) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\varphi_i - \sum_{i=1}^{l}\alpha_i[y_i(w \cdot x + b) - 1 + \varphi_i] - \sum_{i=1}^{l}u_i\varphi_i \tag{5.4}$$

其中，α_i和u_i为整的拉格朗日乘子。函数 L 的极值应满足条件

$$\frac{\partial L}{\partial w} = 0, \frac{\partial L}{\partial b} = 0, \frac{\partial L}{\partial \varphi_i} = 0 \tag{5.5}$$

从而得到

$$\sum_{i=1}^{n}\alpha_i y_i = 0, i = 1, \dots, l \tag{5.6}$$

$$w^* = \sum_{i=1}^{n}\alpha_i y_i x_i, i = 1, \dots, l \tag{5.7}$$

$$C - \alpha_i - u_i = 0, i = 1, \dots, l \tag{5.8}$$

得到：

$$\min F(\alpha) = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{l}\alpha_i \tag{5.9}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, i = 1, \ldots, l \tag{5.10}$$

$$0 \le \alpha_i \le \mathrm{C}, \mathrm{i} = 1, \ldots, \mathrm{n} \tag{5.11}$$

上述问题是一个线性约束的凸二次规划问题，解是唯一的。其中，若解为非零解，则记作支持向量，记作α^*，其判别函数为

$$\mathrm{f(x)} = \mathrm{sgn}(\sum_{SV} \alpha_i^* y_i (x_i \cdot x) + b^*) \tag{5.12}$$

其中，sgn(x)通常情况下为+1 或-1。$b^* = -\frac{1}{2} < w^*, x_r + x_s >$，$x_r$，$x_s$分别来自于每个类别的任一支持向量，即$0 < \alpha_r, \alpha_s < C, y_r = -1, y_s = 1$。

2 核函数

非线性的两类数据由于不再适用于二维空间，只能通过多维空间将其分开。因此，需要引入核函数 K(.,.)进行映射，从而二维空间映射到多维空间变得更加简便。

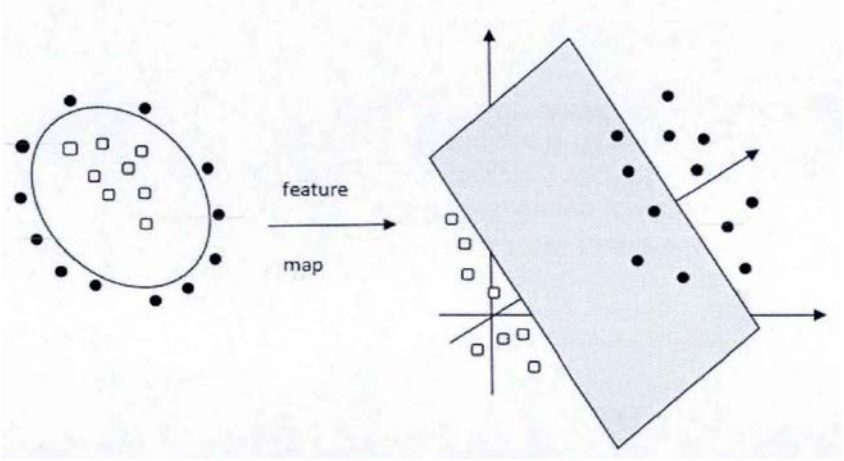


图 5.3　从线性不可分到线性可分

核函数定义：对所有的$\mathrm{x}, \mathrm{z} \in \mathrm{X}$，满足 $\mathrm{k(x,z)} =< \emptyset(x_i) >, \emptyset(\cdot)$为从输入 X 到特征空间 F 的映射。

依据解决不同的问题，选择的核函数也是不一样的。常用的核函数如下：

1.多项式核函数

$$\mathrm{k}(x_1, x_2) = (< x_1, x_2 > + R)^d \tag{5.13}$$

2.线性核函数

$$\mathrm{k}(x_1, x_2) =< x_1, x_2 > \tag{5.14}$$

3.高斯核函数

$$\mathrm{k}(x_1, x_2) = \exp\left\{-\frac{||x_1 - x_2||^2}{2\sigma^2}\right\} \tag{5.15}$$

高斯核函数是使用的较为广泛的核函数之一。由于σ可以进行调节，所以高斯核函数可以根据σ进行调节。若σ比较大，相当于一个低维的子空间；若σ比较小，会导致过拟合。

3 非线性支持向量机

非线性 SVM 无法适用于只用一个最优超平面进行表示，需要将非线性的空间映射到线性空间中。在变换后的空间里利用线性 SVM 进行分类。

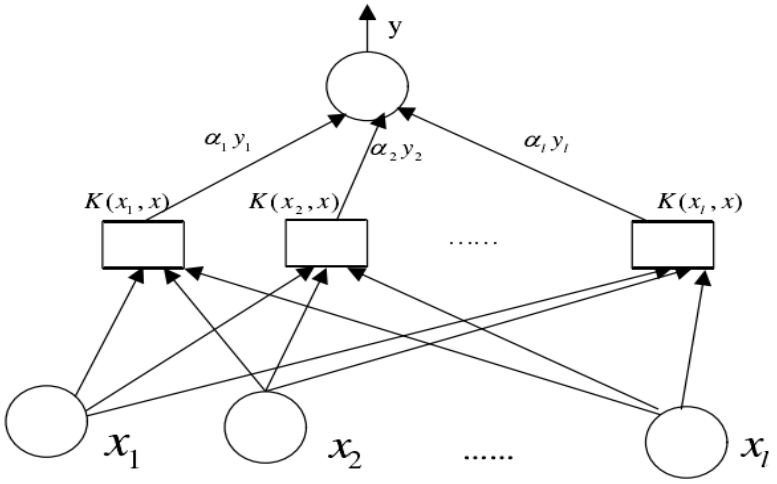


图 5.4　非线性支持向量机示意图

具体算法图下所示：

表 5.2　支持向量机

算法：支持向量机
输入：训练数据集 T={(x1,y1),(x2,y2),...,(xN,yN)}，其中$\mathbf{x_i} \in \boldsymbol{X} = \boldsymbol{R^n}$，$\mathbf{y_i} \in \boldsymbol{Y} = \{-\mathbf{1}, \mathbf{1}\}$,i=1,2,…,N; 输出：分类决策函数 过程： 1.选取适当的核函数 K(x,z)和适当的参数 C，构造并求解最优化问题： $$\min_{\alpha} \frac{\mathbf{1}}{\mathbf{2}} \sum_{i=1}^{N} \sum_{j=1}^{N} \boldsymbol{\alpha_i \alpha_j y_i y_j K}(\boldsymbol{x_i}, \boldsymbol{x_j}) - \sum_{I=1}^{N} \boldsymbol{\alpha_i}$$ $$\boldsymbol{s.t.} \quad \sum_{I=1}^{N} \boldsymbol{\alpha_i y_i} = \mathbf{0}$$ $$\mathbf{0} \leq \boldsymbol{\alpha_i} \leq \mathrm{C}, \quad \mathrm{i}=1,2,\ldots,\mathrm{N}$$ 求得最优解$\boldsymbol{\alpha}^* = (\boldsymbol{\alpha_1^*}, \boldsymbol{\alpha_2^*}, \ldots, \boldsymbol{\alpha_N^*})^{\boldsymbol{T}}$

续表 5.2　支持向量机

2.选择$\boldsymbol{\alpha}^*$的一个正分量$\mathbf{0} < \boldsymbol{\alpha}_j^* < \boldsymbol{C}$，计算

$$\boldsymbol{b}^* = \boldsymbol{y}_j - \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{\alpha}_i^* \boldsymbol{y}_i \boldsymbol{K}(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

3.构造决策函数：

$$\mathbf{f(x)} = \mathbf{sign}(\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{\alpha}_i^* \boldsymbol{y}_i \boldsymbol{K}(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + \boldsymbol{b}^*)$$

当是正定核函数时，XX 是凸二次规划问题，解是存在的。

5.2.3　集成学习算法

集成学习利用多个学习器来完成训练任务。通常情况下，可以获得比单一学习器显著优越的泛化性能。图 5.5 为集成学习的一般结构：第一步，生成一组“个体学习器”；第二步，采用某种策略将第一步生成的学习器结合。

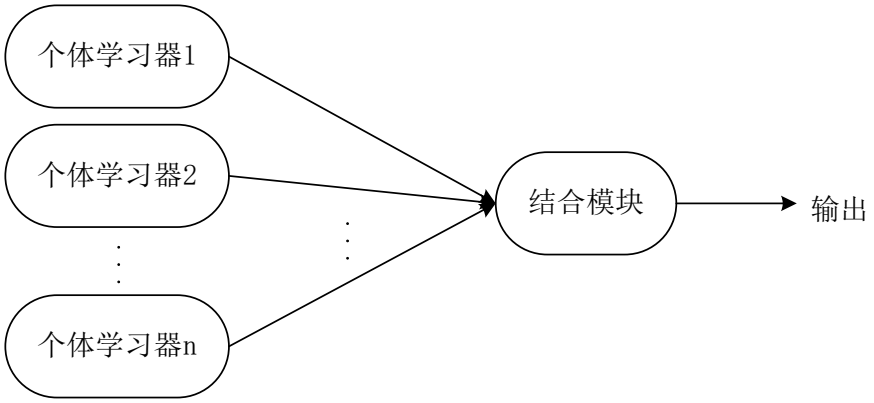


图 5.5　集成学习示意图

在分类问题中，个体学习器也称为基学习器有时也被成为弱学习器，由于生成强分类器需要比较高的要求，因此生成弱分类器相对简单。随机梯度提升树（Gradient Boosting）是在把弱学习器提升为强学习器的思想下的监督训练模型优化方法。随机梯度提升树的每一次的训练可以减少上一次的残差，为了能够不断地减小残差，因此需要在减小残差的梯度方向重新训练一个新的模型。首先需要将集成模型进行拆分，拆分成若干个弱分类器。将这些弱分类器在梯度的方向上降低残差，使原来的弱分类器训练为一个强分类器，检测的效果相比于前者而言更加精准。

GBDT 可表示成决策树的加法模型：

$$f_M(\mathrm{x}) = \sum_{m=1}^{M} T(x; \vartheta_m) \tag{5.16}$$

其中，$T(x;\vartheta_m)$表示决策树，ϑ_m表示为一棵树的参数，M 为决策树的数量。

在决策树的构造中，每一棵树都学习上一课树的残差，换句话说，当前决策树的产生的结果就是上一棵树的结果和残差的和。例如：确定第一步的提升树$f_0(x)=0$,那么第 m 步的模型为$f_m(x)=f_{m-1}(x)+T(x;\vartheta_m)$。

其中，$f_{m-1}(x)$为当前的模型，通过经验风险最小化的原则确定下一颗决策树的参数ϑ_m。

$$\widehat{\vartheta_m}=arg\min_{\vartheta_m}\sum_{i=1}^{N}L(y_i,f_{m-1}(x_i)+T(x_i;\vartheta_m)) \tag{5.17}$$

对应的损失函数即均方误差损失：

$$\mathrm{L}\big(\mathrm{y},\mathrm{f}(\mathrm{x})\big)=\big(y-f(x)\big)^2 \tag{5.18}$$

具体的算法流程如表 5.3 所示：

表 5.3　梯度提升树提升算法

算法 梯度提升树算法流程
输入：训练数据集 S
输出：梯度提升树模型
过程
1.初始化 $\boldsymbol{f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c)}$
2.for m=1,2,3,…,M do:
3. $\boldsymbol{\tilde{y}_i=-[\frac{\partial L(y_i,F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}}$, i=1,…,N
4. $\boldsymbol{\alpha_m=arg\min_{\alpha,\beta}\sum_{i=1}^{N}[\tilde{y}_i-\beta h(x_i;\alpha)]^2}$
5. $\boldsymbol{c_m=arg\min_c\sum_{i=1}^{N}L(y_i,F_{m-1}(x_i)+ch(x_i;\alpha_m))}$
6.$\boldsymbol{F_m(x)=F_{m-1}(x)+c_mh(x;\alpha_m)}$
7.end for
8.end Algorithm

5.2.4　基于 AETA 数据的震级预测模型评估指标及选择

混淆矩阵用于把实际的样本值和模型的预测值进行联表分析，本文应用混淆矩阵作为模型评估的指标。

二分类模型的性能评估，如表 5.4 所示：

表 5.4　分类模型性能评估

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

其中 Positive/Negative 对应于模型预测的分类结果，而 True/False 表示分类是否正确。

各个数据的含义如表 5.5 所示：

表 5.5 指标数据含义

	含义
TN	模型将负类预测为负类的样本数
FN	模型将正类预测为负类的样本数
FP	模型将负类预测为正类的样本数
TP	模型将正类预测为正类的样本数

评价标准定义为：

（1） 正确率(accuracy)：表示分类器对整个样本的判别能力，即判定正确的数量占总样本书的比例，表达式为：

$$\text{accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \tag{5.19}$$

（2） 准确率(precision)：反应了被正确判定的正类的数量占模型预测正样本总数之比，衡量模型预测正样本的准确性，表达式为：

$$\text{precision} = \frac{TP}{TP+FP} \tag{5.20}$$

（3） 查全率(recall):反映了被正确判定的正类数量占实际正样本数之比，可以衡量模型预测正样本的可信性，表达式为：

$$\text{recall} = \frac{TP}{TP+FN} \tag{5.21}$$

（4） F1-score:准确率与查全率的加权调和平均，表达式为：

$$\text{F1} = \frac{2*precision*recall}{precision+recall} \tag{5.22}$$

本文用准确率，查全率和 F1-score 和 ROC 曲线的面积（Aear Under Curve，AUC）来评价预测效果。其中，AUC 的值介于 0-1。AUC 可直观评价分类器的好坏，当值接近 0.5 时分类器的效果最差，效果与随机预测等同。当 AUC 的值大于 0.5，越接近于 1 时性能越好。

本文建立一个小样本集对所研究的三种预测模型进行了分析比较。其中实验数据采用对 5.4.1 节中的数据进行 5:1 抽取，样本空间的构建方式不变，实验目标是预测研究区域内未来 15 天是否发生地震，为一个二分类问题。对本文研究的 CART 决策树和 SVM 以及集成学习算法利用上述指标进行性能对比，以选择出可用于 AETA 的地震震级预测的模型。

得到的指标如表 5.6 所示：

表 5.6　三种模型的性能比较

	CART	SVM	GBDT
AUC	0.528	0.562	0.685

从表中可以得到，梯度提升树的 AUC 值为 0.685，优于 CART 的 0.528 和 SVM 的 0.562，说明梯度提升树的分类能力较好，预测效果较好。因此，本文选择梯度提升树对地震震级进行预测研究。

5.3　基于 AETA 数据和多元线性回归的地点预测模型的研究

多元线性回归表示为因变量和多个自变量之间的线性关系，数学模型为：

$$y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots+\beta_nX_n+\varepsilon \tag{5.23}$$

其中，β_0 为回归常数，β_1，$\cdots\cdots$，β_n 为偏回归系数。

从上式可以看出，y 总共分为两部分。一部分是前面的 $\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots+\beta_nX_n$，表示的 y 的线性部分；第二部分是 ε，ε 由于不可控制的随机因素引起的误差，表示的是 y 的非线性部分。

则多元线性回归模型的流程图如图 5.6 所示：

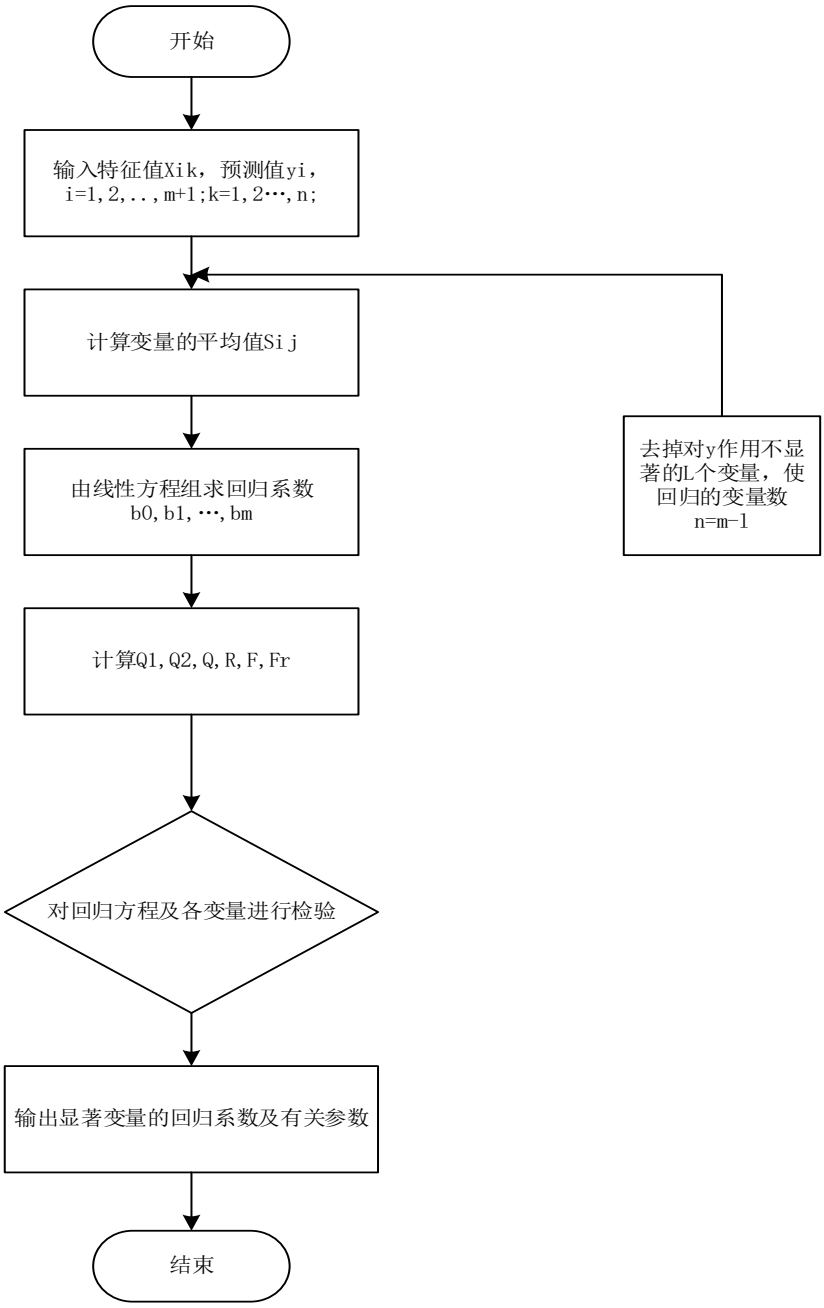


图 5.6　多元线性回归流程图

5.4　基于 AETA 数据的地震预测模型的建立与评估

5.4.1　实验数据的选取

1976 年 7 月 28 日，河北唐山发生了 7.8 大地震；2008 年 5 月 12 日，四川汶川发生了 8 级地震；2017 年 8 月 8 日，四川九寨沟发生了 7 级地震；这些大地震都造成了极其巨大的人员伤亡和财产损失。这些都使人深省，如果能够实现地震预测，有关部门

能尽早采取措施，人员进行梳理，将会在很大程度上降低人员的伤亡和财产的损失。因此本文在进行地震区域选择考虑地震发生密集的区域，以及曾经发生历史大震的区域。在进行震级的选取，本文希望能够预测对人们人身和财产安全受到影响的大震。

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。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。

经统计，每年发生 4 级以上地震总共本文的目的进行地震预测以减少由于地震造成的人员伤亡。在研究地震预测问题上，由于 4 级以及 4 级以下的地震发生比较频繁，4 级 4 级以上的地震对人身安全以及财产安全更具有威胁性，因此本文只针对 4 级及 4 级以上的地震。

同时地区的选择也十分重要。如果选择 4 级地震发生不是十分频繁的地区，数据量的不足导致研究成果可信度会降低。中国的四川、云南、河北等地区有一些区域的地震发生较为频繁，此区域更具有研究意义。本文地震的预测目标区域选择为地震发生较为频繁的区域 1：[100,103,25,28];区域 2：[103,106,31,34]两个区域。数据选取自设备布设 2017 年 6 月 1 日至 2019 年 3 月 1 日的数据。

区域 1 的地理位置以及台站分布和震源图，如图 5.7 所示：

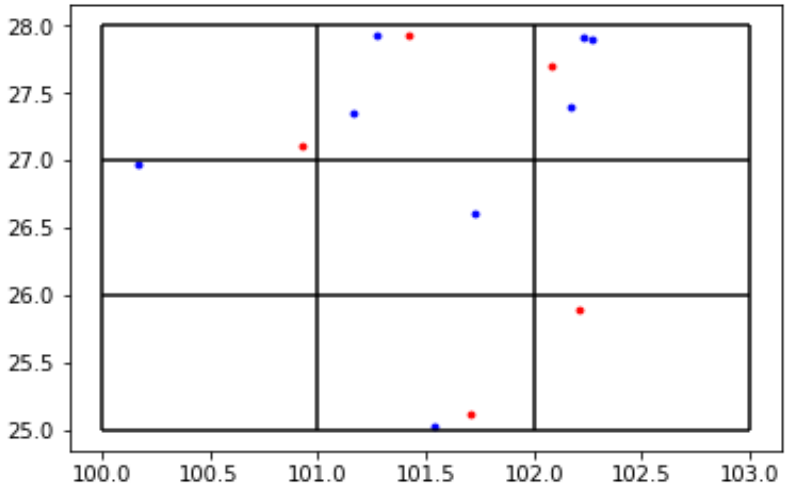


图 5.7　区域 1 台站分布及震源图

区域 2 的地理位置以及台站分布和震源图，如图 5.8 所示：

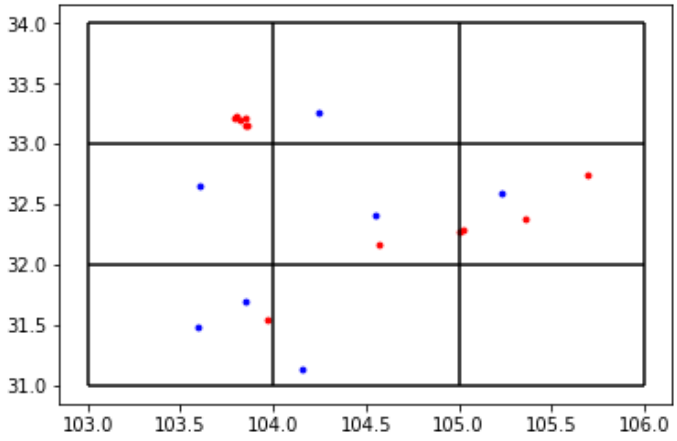


图 5.8　区域 2 台站分布图及震源图

区域 1 的地震信息，如表 5.7 所示：

表 5.7　区域 1 的地震信息

序号	时间	震级	纬度	经度	深度	位置
1	2017-07-02 14:34:56	4.1	25.12	101.71	5.0	云南楚雄州禄丰县
2	2017-09-12 19:26:40	4.4	27.93	101.42	13.0	四川凉山州木里县
3	2018-10-17 13:29:18	4.5	25.89	102.21	11.0	云南楚雄州武定县
4	2018-10-31 16:29:55	5.1	27.70	102.08	19.0	四川凉山州西昌市
5	2019-02-02 05:26:06	4.1	27.11	100.93	15.0	云南丽江市宁蒗县

区域 2 的地震信息，如表 5.8 所示；

表 5.8　区域 2 的地震信息

序号	时间	震级	经度	经度	深度	位置
1	2017-07-17 06:55:59	4.9	32.38	105.36	21.0	四川广元市青川县
2	2017-08-08 21:19:46	7.0	33.20	103.82	20.0	四川阿坝州九寨沟县

续表 5.8　区域 2 的地震信息

3	2017-09-30 14:14:37	5.4	32.27	105.00	13.0	四川广元市青川县
4	2017-10-06 18:25:34	4.0	33.23	103.80	20.0	四川阿坝州九寨沟县
5	2017-11-07 05:31:08	4.5	33.21	103.79	16.0	四川阿坝州九寨沟县
6	2017-11-10 11:03:14	4.2	31.54	103.97	19.0	四川德阳市绵竹市
7	2018-02-18 11:44:11	4.4	32.29	105.02	19.0	四川广元市青川县
8	2018-06-29 08:42:18	4.0	32.17	104.57	18.0	四川绵阳市平武县
9	2018-09-12 19:06:34	5.3	32.75	105.69	11.0	陕西汉中市宁强县

在这两个区域内，[100,103,25,28]具体的台站如表 5.9 所示：

表 5.9　区域 1 的台站列表

序号	台站名	台站号	经度	纬度	安装时间
1	楚雄山洞	29	101.54	25.03	2016-12-30
2	西昌气象局	32	102.27	27.90	2016-12-19
3	德昌防震减灾局	41	102.17	27.40	2017-04-26
4	丽江山洞	55	100.17	26.97	2017-01-27
5	盐源县盐塘乡	78	100.17	27.35	2017-04-22
6	木里县防震减灾局	101	101.27	27.93	2017-04-23
7	马兰山地震台	109	101.73	26.60	2017-04-25
8	西昌小庙山洞	132	102.73	27.91	2017-06-18

[103,106,31,34]具体的台站如表 5.10 所示：

表 5.10　区域 2 的台站列表

序号	台站名	台站号	经度	纬度	安装时间
1	青川县防震减灾局	43	105.23	32.59	2017-06-10
2	茂县测点	90	103.85	31.69	2017-06-17
3	汶川防震减灾局	91	103.59	31.48	2017-06-17

续表 5.10　区域 2 的台站列表

4	什邡市防震减灾局	99	104.16	31.13	2017-06-09
5	平武县防震减灾局	116	104.55	32.41	2017-03-18
6	九寨沟防震减灾局	121	104.25	33.26	2017-06-13
7	松潘地震台	129	103.60	32.65	2017-06-16

从表 5.6 和表 5.7 中可知，震级预测区域内在 2017 年 6 月 1 日到 2019 年 3 月 1 日之间符合本文研究的 4 级以上的地震在区域中只有 5 个，在区域二中只有 9 个。对于机器学习任务训练显然是太少了。

因此，本文预测未来 15 天是否发生 4 级以上的地震，采用较小的步长增大样本。若采用步长为 1 天，那么震前 15 天到震前 1 天则总可以产生 15 个样本。如此，区域一的 5 个地震则可以产生 75 个正样本，区域二的 9 个地震则可以产生 135 个正样本。其中，训练集和验证集的划分，由于地震事件有限，因此本文选区域 1 的最新发生的两个地震作为验证集的正样本，选区域 2 最新发生的 3 个地震作为验证的正样本。

5.4.2　实验结果与分析

实验参与表 4.6 中的实验参数相同。在 5.4.1 小节中的数据集中，首先对地震的震级进行预测，在验证集上模型的地震预测结果如图 5.9 和表 5.11 所示。

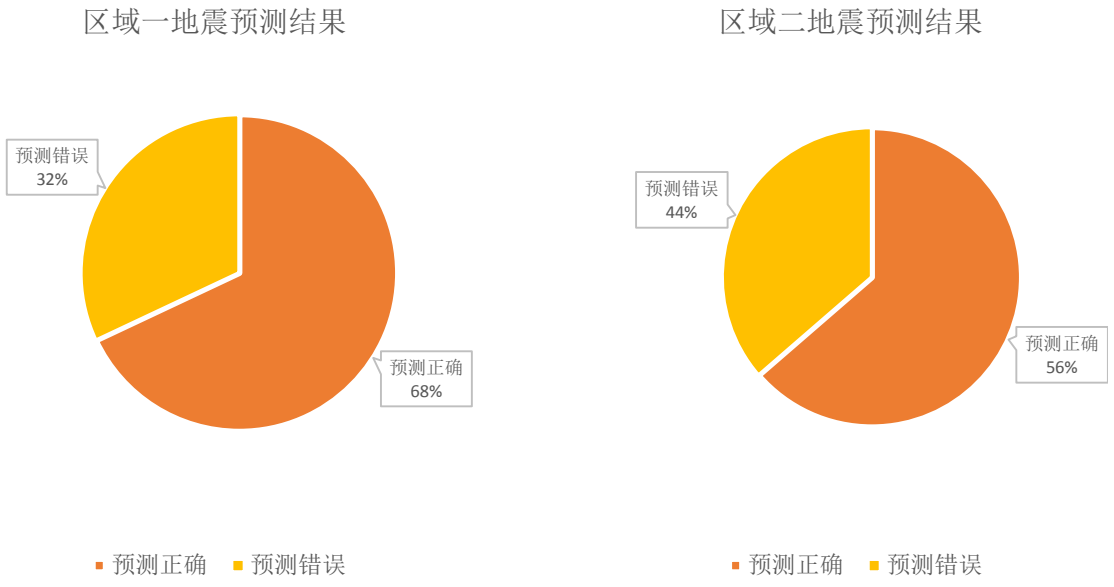


图 5.9 梯度提升树对地震预测结果

表 5.11　地震震级预测结果对比

指标	区域 1	区域 2
Precision	0.68	0.56
Recall	0.57	0.62
F1-score	0.62	0.59

不过，因为在验证集上区域 1 中只有 2 个地震，区域 2 有 3 个地震，区域内震例数量太少，仍然需要进一步数据积累来证明模型在所选区域内的地震事件时间和震级的预测效果。

本文利用 5.3 节中的回归模型对地震的震中经纬度进行预测，在验证集中的预测结果如表 5.12 和表 5.13 所示：

表 5.12　区域 1 地震震中预测结果

序号	时间	震级	实际震中	预测震中
1	2018-10-31	5.1	(102.08,27.70)	(102.71,26.84)
2	2019-02-02	4.1	(100.93,27.11)	(100.68,27.63)

表 5.13　区域 2 地震震中预测结果

序号	时间	震级	实际震中	预测震中
1	2018-02-18	4.4	(105.02,32.29)	(104.11,33.34)
2	2018-06-29	4.0	(104.57,32.17)	(105.68,31.28)
3	2018-09-12	5.3	(105.69,32.75)	(103.92,32.67)

区域 1 的验证集的震中经度均方根误差(RMSE)为 0.48，纬度均方根误差(RMSE)为 0.71；区域 2 的验证集中经度均方根误差(RMSE)为 1.31，纬度均方根误差(RMSE)为 0.80。根据经纬度的距离换算公式，在区域 1 和区域 2 中，经度 1 度约为 111Km，纬度 1 度约为 97Km，则模型在区域 1 中的震中经度预测平均误差为 53.28Km，纬度预测平均误差为 68.87Km。在区域 2 中的震中经度预测平均误差为 145.41Km，纬度预测平均误差为 77.60Km。由于验证集上的地震数量仅有 5 个，数量不充足，仍然需要进一步的数据积累来证明模型在所选区域内的地震震中的预测效果。

本研究不仅基于数据本身的特点进行特征提取，还利用希尔伯特黄变换对 AETA 数据进行震前瞬时能量分析。然后对所选择出的对地震贡献度大的特征利用机器学习算法进行地震的震级预测和地点预测，相比较与传统的仅基于前兆特征进行地震预测以及现有的基于算法的地震预测相比较，如表 5.14 所示：

表 5.14　本文工作与其他相关工作对比表

	前兆观测量	数据窗口	研究区域	震级预测结果	震中预测结果
ELF 电磁信号	ELF 电磁信号	1 年	陇南	-	一次相距 107km
GMM 模型	电离层	6 个月	日本部分区域	准确率 0.63	-
HMM 模型	电离层	6 个月	日本部分区域	准确率 0.55	-
危险理论	历史地震	5~15 天		—	准确率 0.52
本文	电磁和地声	20 个月	云南部分区域	准确率 0.68	RMSE:经度为 0.48，纬度为 0.71；
本文	电磁和地声	20 个月	四川部分区域	准确率 0.56	RMSE:经度为 1.31，纬度为 0.80

5.5　本章小结

本章研究了地震预测模型的建模方法，并进行了相关的数据实验。首先介绍了地震三要素的预测结构，以及地震三要素的选择。然后研究了 cart 决策树，支持向量机，stacking 集成算法梯度提升树并对比了三种算法的地震震级的预测效果。利用多元线性回归进行地震地点的预测。并且在东经 100°~103°，北纬 25°~28°以及东经 103°~106°，北纬 31°~34°两个区域，2017 年 6 月 1 日至 2019 年 3 月 1 日期间的数据集上进行了数据实验，得到模型在区域 1 验证集中地震事件的准确率 0.68，查全率为 0.57；在区域 2 验证集中地震时间的准确率为 0.56，查全率为 0.62。区域 1 验证集中震中经纬度预测均方误差分别为 0.48 和 0.71。区域 2 验证集中震中经纬度均方误差分别为 1.31 和 0.80。

第六章　总结与展望

6.1　总结

地震预测在现代研究领域中是一个集聚热点、难点的课题。本文基于 AETA 监测数据，进行地震三要素预测的研究。从电磁数据和地声数据的均值、振铃计数、峰值频率三个分量为基础，基于波形特点及希尔伯特黄变换进行了特征的提取，并基于随机森林对所提取的 66 个特征进行选择出 20 个特征，最后利用梯度提升树进行震级预测，利用多元线性回归进行地震地点的预测，取得了一定的效果。

本文完成的工作如下：

1.基于 AETA 数据的预处理

从 AETA 系统监测的数据特点出发，由于在实际监测中不可避免地会出现断电、断网等情况，因此数据会出现缺失的情况。对于断网导致的缺失数据，利用线性插值的方法进行补全，对于断电导致的缺失数据，采用设置阈值方法进行判断并通过线性插值方法进行补全。使数据变得完整，为之后的数据分析工作奠定了基础。

2.基于 AETA 数据进行特征提取

本文基于波形特点及信号变换两方面对电磁数据和地声数据进行特征提取。首先基于电磁数据出现日周期的波形特点，提取出了 7 种波形，并对台站进行波形相似性分析，提取波形出现的次数，共得到 21 种特征。其次，基于地声数据在震前会出现尖峰的特点，对波峰进行提取，共得到 9 种特征。最后，基于信号变换方法对电磁数据和地声数据进行特征提取。本文提出了基于希尔伯特-黄变换的能量分析法，首先将电磁和地声信号利用 EMD 进行分解，选取 imf1 分量，对其进行希尔伯特变换得到震前波形的瞬时能量变化，共得到 36 种特征。

3.特征权重评估

基于第三章总共提取达 66 种的特征，总共提取的 66 种特征进行分析，发现其中包含一些冗余特征。这些冗余的特征会导致分类器计算开销太大，性能较差，导致建立模型产生困难，因此本文进行特征权重评估，以得到对地震预测贡献度高的特征。首先对 relief-F 算法，LVW 算法，以及随机森林进行研究分析并对比其优缺点。最后利用随机森林对所提取的特征集进行了权重评估，筛选出了对地震预测贡献度最大 energy_peak_max、ring_15、energy_sum 等 20 个特征。

4.地震预测模型研究

以地震时间、震级、震中三要素的预测为目标，利用筛选后的 20 个特征，使用梯度提升树建立了地震预测模型，并且在东经 100°~103°，北纬 25°~28°以及东经 103°~106°，北纬 31°~34°两个区域，2017 年 6 月 1 日至 2019 年 3 月 1 日期间的数据集上进行了数据实验，得到模型在区域 1 验证集中地震事件的准确率 0.68，查全率为 0.57；在区域 2 验证集中地震时间的准确率为 0.56，查全率为 0.62。区域 1 验证集中震中经纬度预测均方误差分别为 0.48 和 0.71。区域 2 验证集中震中经纬度均方误差分别为 1.31 和 0.80。实验结果表明，本文的工作对于解决地震预测问题有一定意义。

6.2 展望

本研究以 AETA 系统监测的电磁数据和地声数据出发，进行地震三要素的预测研究虽然取得了一定的效果，但是在未来的的研究中还有待进一步的完善。主要有以下几点：

1.增加震例的数量。本文此次研究的震例仅为两个区域的四级以上的地震，数量有限，所含的震例还是不足，还是需要积累震例进行进一步的研究，达到更精准的效果。

2.适用范围的扩大。本文此次研究仅仅两个历史大震区，并没有对所有区域进行研究，还需要扩大研究区域，使地震预测能够达到普适的效果，并应用到实际中，以期减轻人员伤亡以及财产损失。

3.数据的深入挖掘。本研究仅对电磁数据以及地声数据本身所含的特点进行提取与总结以及基于希尔伯特黄变换的能量分析，但是由 AETA 系统实际监测的数据价值很高，数据本身所含的信息量是巨大的，还需要进一步的挖掘对地震预测更有用的信息，以对之后进行地震预测做出贡献。

参考文献

[1] 赵克常. 地震概论[M]. 北京：北京大学出版社, 2012.

[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] 国家地震局科技监测司. 地震观测技术[M]. 地震出版社, 1995.

[4] 袁桂平, 李鸿宇, 张贵霞等. 地磁垂直分量 Z 日变幅逐日比及其与磁暴和地震的关系[J]. 地震, 2018, 38(1):139-146.

[5] Thomas J N , Love J J , Johnston M J S , et al. On the reported magnetic precursor of the 1993 guam earthquake[J]. Translated World Seismology, 2009, 36(16):16301.

[6] Alcay S . Anamolous ionospheric TEC variations prior to the Indonesian earthquake (M 7.1) of November 15, 2014[J]. Geomagnetism and Aeronomy, 2017, 57(3):301-307.

[7] 孔令昌, 王桂清. 大地微电流异常用于地震预报的可能性[J]. 地震地磁观测与研究, 2009, 30(5):71-77.

[8] 杨亦春, 郭泉, 吕君等. 大地震前出现的异常次声波观测研究[J]. 物理学报,2014,63(13):224-237.

[9] 吕君, 郭泉, 冯浩楠等. 北京地震前的异常次声波[J]. 地球物理学报, 2012, 55(10):3379-3385.

[10] Nikonov A A . Contribution to earthquake prediction by the data of recent crustal movement anomalies[J]. Tectonophysics, 1979, 52(1-4):644-645.

[11] Wanju B , Juzhong C , Jianfeng S , et al. ACCURACY OF GEODETIC DATA AND ITS ROLE IN RESEARCHES ON CRUSTAL DEFORMATION AND EARTHQUAKE PREDICTION[J]. Journal of Geodesy & Geodynamics, 2011, 31(1):44-48.

[12] 闫相相, 单新建, 曹晋滨等. 日本 M_W 9.0 级特大地震前电离层扰动初步分析[J]. 地球物理学进展, 2013, 28(1):155-164.

[13] 刘祎, 周晨, 赵正予等. 基于 LAIC 电场渗透和 SAMI2 模拟的地震-电离层扰动现象研究[J]. 地震, 2018, 38(1):74-83.

[14] Zuji Q , Ainai M , Zuoxun Z . A study of the method of satellite thermal infrared earthquake prediction in imminence[J]. Earth Science Frontiers, 2010, 17(5):254-262.

[15] 张璇, 张元生, 郭晓等. 尼泊尔 8.1 级地震卫星热红外异常解析[J]. 地学前缘,2017,24(02):227-233.

[16] Liu S , Yang D , Ma B , et al. On the features and mechanism of satellite infrared anomaly before earthquakes in Taiwan Region[C]// IEEE International Geoscience & Remote Sensing Symposium, IGARSS 2007, July 23-28, 2007, Barcelona, Spain, Proceedings. IEEE, 2007.

[17] 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.

[18] 赵永红, 谢雨晴, 王航等. 地震预测方法Ⅴ:地下流体方法[J]. 地球物理学进展, 2017(04):123-131.

[19] MILNE, JOHN. Seismology in Japan[J]. Nature, 1880, 22(557):208-208.

[20] Cornell C A. Engineering seismic risk analysis[J]. Bulletin of the Seismological Society of America, 1968, 58(11 Suppl 1):S183-S188.

[21] Papazachos B C , Karakaisis G F , Papazachos C B , et al. Perspectives for earthquake prediction in the Mediterranean and contribution of geological observations[J]. Geological Society, London, Special Publications, 2006, 260(1):689-707.

[22] 尹祥础. 地震预测新途径的探索[J]. 中国地震, 1987(1):3-10.

[23] 张浪平, 尹祥础, 梁乃刚. 加卸载响应比在伊朗地区地震活动性研究中的应用[J]. 中国地震, 2006, 22(4):356-363.

[24] 秦四清, 徐锡伟, 胡平等. 孕震断层的多锁固段脆性破裂机制与地震预测新方法的探索[J]. 地球物理学报, 2010, 53(4):1001-1014.

[25] 秦四清, 薛雷, 王媛媛等. 对孕震断层多锁固段脆性破裂理论的进一步验证及有关科学问题的讨论[J]. 地球物理学进展, 2010, 25(3):749-758.

[26] 陈棋福, 石耀霖. 基于遗传算法的分类体系在地震预报中的应用探索[J]. 地球物理学报, 1997, 40(4):539-549.

[27] 李莹甄, 王海涛, 龙海英等. 基于遗传算法的地震短期综合预报分类系统研究[J]. 地震工程学报, 2002, 24(4):295-302.

[28] 李莹甄, 王海涛, 龙海英. 基于遗传算法地震短期综合预报分类系统在天山地震带的应用[J]. 内陆地震, 2004, 18(1):1-13.

[29] 李荣峰. 福建及其周边地区地震活动人工神经网络模型的构建[J]. 应用海洋学学报, 2000, 19(1):107-112.

[30] Rundle J B , Klein W , Turcotte D L , et al. Precursory Seismic Activation and Critical-point Phenomena[J]. Pure and Applied Geophysics, 2000, 157(11-12):2165-2182.

[31] Holliday J R , Rundle J B , Tiampo K F , et al. Systematic Procedural and Sensitivity Analysis of the Pattern Informatics Method for Forecasting Large (M > 5) Earthquake Events in Southern California[J]. Pure and Applied Geophysics, 2006, 163(11-12):2433-2454.

[32] Asim K M , Adnan I , Talat I , et al. Earthquake prediction model using support vector regressor and hybrid neural networks[J]. PLOS ONE, 2018, 13(7):e0199004-.

[33] Nanjo K Z, Holliday J R, Chen C C, et al. Application of a modified pattern informatics method to forecasting the locations of future large earthquakes in the central Japan[J]. Tectonophysics, 2006, 424(3):351-366.

[34] 张炜, 阎立璋, 申春生等. 水文地球化学地震前兆观测与新灵敏组分的探索[J]. 地震, 1987(5):58.

[35] 张桂清. 全球性地震活动与太阳活动的关系[J]. 地震学报, 1998(4):427-431.

[36] 丁鉴海, 申旭辉, 潘威炎等. 地震电磁前兆研究进展[J]. 电波科学学报, 2006, 21(5):791-801.

[37] Zeng X , Lin Y , Xu C , et al. Turning changes in evolution of geomagnetic field and infrastructural analysis of earthquake prediction[J]. Kybernetes, 2001, 30(4):365-377.

[38] Suratgar A A , Setoudeh F , Salemi A H , et al. Magnitude of Earthquake Prediction Using Neural Network[C]// Fourth International Conference on Natural Computation. IEEE Computer Society, 2008.

[39] Hayakawa M . Current Status of Seismo Electromagnetics as Short-term Earthquake Prediction[J]. Ieej Transactions on Sensors & Micromachines, 2010, 130(7):431-434.

[40] 刘君, 安张辉, 范莹莹等. 芦山 MS7.0 与岷县漳县 MS6.6 地震前电磁扰动异常变化[J]. 地震, 2015, 35(4).

[41] 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.

[42] 李军辉, 何康, 郑海刚等. 安徽阜阳 4.3 级地震前电磁扰动异常分析[J]. 四川地震, 2018(3):29-32.

[43] 张建国, 刘晓灿, 姚丽等. 汶川 8.0 级大地震前电磁扰动异常变化特征初步研究[J]. 地震地磁观测与研究, 2010, 31(5).

[44] 丁跃军. 电磁扰动映震分析[C]// 中国地震学会地震电磁学专业委员会、中国地震学会空间对地观测专业委员会会暨学术研讨会会议摘要. 2010.

[45] Kopytenko Y A, Ismagilov V S, Schekotov A, et al. Peculiarities of ULF electromagnetic disturbances before strong earthquakes in seismic active zone of Kamchatka peninsula[C]// Agu Fall Meeting. 2006.

[46] 蒋淳, 田山, 陈化然等. 地震综合预测物元模型及其应用[J]. 地震学报, 2000, 22(4).

[47] Keilis-Borok V , Shebalin P , Gabrielov A , et al. Reverse Detection of Short-Term Earthquake Precursors[J]. Physics of the Earth & Planetary Interiors, 2003, 145(1):75-85.

[48] Bowman D D , King G C P . Accelerating seismicity and stress accumulation before large earthquakes[J]. Geophysical Research Letters, 2001, 28(21):4039–4042.

[49] Gelfand I M, Guberman S A, Keilis-Borok V I, et al. 'Pattern Recognition Applied to Earthquake Epicenters in California'[J]. Physics of the Earth & Planetary Interiors, 1976, 11(3):227-283.

[50] Shebalin P , Narteau C , Holschneider M . From Alarm-Based to Rate-Based Earthquake Forecast Models[J]. Bulletin of the Seismological Society of America, 2012, 102(1):64-72.

[51] Shebalin P N , Narteau, Clément, Zechar J , et al. Combining earthquake forecasts using differential probability gains[J]. Earth, Planets and Space, 2014, 66(1):37.

[52] 薄万举, 吴翼麟. 信息合成方法及其应用研究[J]. 大地测量与地球动力学, 1995(3):84-88.

[53] 韩天锡, 蒋淳, 魏雪丽等. 多元统计组合模型在地震综合预报中的应用[J]. 地震学报, 2004, 26(5):523-528.

[54] 王海涛, 曲延军, 和锐. 基于多种地震前兆异常的综合异常指数研究[J]. 内陆地震, 2002, 16(4):302-305.

[55] 张祖胜, 杨国华. 地壳垂直形变速率梯度,断层形变速率变化与强震危险区研究[J]. 中国地震, 1996(4):347-357.

[56] 楼海, 王椿镛. 川滇地区重力异常的小波分解与解释[J]. 地震学报, 2005, 27(5):515-523.

[57] 王吉易, 郑云贞, 刘允清等. 水氡灵敏点地震水文地球化学条件探讨[J]. 地震, 1988(1):48-51.

[58] 赵永红. 活动断裂带附近地下水中的氢同位素变化与地震关系研究[J]. 岩石学报, 2011, 27(6):1909-1915.

[59] 袁洁浩, 顾左文, 陈斌等. 美国的震磁观测与研究[J]. 地震研究, 2014(1):163-169.

[60] 尹亮,杨立明.宽频带数字资料低频波在大震前的短临前兆信息研究[J].西北地震学报,2010,32(01):82-87

[61] 洪星, 杨贵, 林仙坎等. 数字台网观测资料中 sPn 震相的测定与应用[J]. 福建地震, 2004(3):14-17.

[62] Leonard R S, Jr R A B. Observation of ionospheric disturbances following the Alaska earthquake[J]. Journal of Geophysical Research, 1965, 70(5):1250-1253.

[63] Parrot M . High-frequency seismo-electromagnetic effects[J]. Phys. Earth Plant. Int. 1993, 77(1):65-83.

[64] Pulinets S A, Legen'Ka A D, Gaivoronskaya T V, et al. Main phenomenological features of ionospheric precursors of strong earthquakes[J]. Journal of Atmospheric and Solar-Terrestrial Physics, 2003, 65(16):1337-1347.

[65] 丁丹,倪四道,田晓峰,敬少群.地震相关的声音现象研究进展[J].华南地震,2010,30(02):46-53.

[66] Gornyi V I, Sal'Man A G, Tronin A A, et al. Outgoing infrared radiation of the earth as an indicator of seismic activity[J]. Doklady Akademii Nauk Sssr, 1988, 301:67-69.

[67] Shebalin P . Increased correlation range of seismicity before large events manifested by earthquake chains[J]. Translated World Seismology, 2007, 424(3):335-349.

[68] 李青梅, 张元生, 吕俊强等. 2014 年 10 月 7 日云南景谷 M_S6.6 地震热红外异常[J]. 地震工程学报, 2015(4):1007-1012.

[69] 郭晓, 张元生, 钟美娇等. 提取地震热异常信息的功率谱相对变化法及震例分析[J]. 地球物理学报, 2010, 53(11).

[70] 张璇, 张元生, 魏从信等. 云南彝良 5.7 级地震前卫星热红外异常[J]. 地震工程学报, 2013(1):171-176.

[71] Mao J, Zhu Y. Progress in the Application of Ground Gravity Observation Data in Earthquake Prediction[J]. Advances in Earth Science, 2018.

[72] Sarlis N V, Varotsos P A, Skordas E S, et al. Seismic electric signals in seismic prone areas[J]. Earthquake Science, 2018, v.31(01):45-52.

[73] 张轶鹏. 地震电磁疑似异常特征提取与地震相关性分析[D]. 2010.

[74] 周挚, 山秀明, 张立等. 基于 HHT 提取昆明、下关重力固体潮的地震前兆信息[J]. 地球物理學報, 2008, 51(3):836-844.

[75] 张璇, 田秀丰, 张俏丽. 活动断裂的红外地震异常特征提取——汶川地震为例[J]. 国际地震动态, 2018, No.476(08):17-18.

[76] Wang Q , Guo Y , Yu L , et al. Earthquake Prediction based on Spatio-Temporal Data Mining: An LSTM Network Approach[J]. IEEE Transactions on Emerging Topics in Computing, 2017:1-1.

[77] Allamehzadeh M , Durudi S , Mahshadnia L . Pattern recognition of seismogenic nodes using Kohonen selforganizing map: example in west and south west of Alborz region in Iran[J]. Earthquake Science, 2017(03):33-43.

[78] Asim K M , Adnan I , Talat I , et al. Earthquake prediction model using support vector regressor and hybrid neural networks[J]. PLOS ONE, 2018, 13(7):e0199004-.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] Huang N E, Wu Z. A review on Hilbert‐Huang transform: Method and its applications to geophysical studies[J]. Reviews of Geophysics, 2008, 46(2):-.

[86] 韩松, 何利铨, 孙斌等. 基于希尔伯特-黄变换的电力系统低频振荡的非线性非平稳分析及其应用[J]. 电网技术, 2008, 32(4):56-60

[87] 杨培杰, 印兴耀, 张广智. 希尔伯特-黄变换地震信号时频分析与属性提取[J]. 地球物理学进展, 2007, 22(5):1585-1590.

[88] 张小飞, 陶凌, 邓娟等. 基于希尔伯特-黄变换的白细胞信号分析[J]. 中国生物医学工程学报, 2014, 33(1):57-62.

[89] M. Dash H L. Feature Selection for Classification[J]. Intelligent Data Analysis, 1997, 1(3):131-156.

[90] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF[C]// European Conference on Machine Learning on Machine Learning. 1994.