版权声明

摘要

大地震是一种具有高度破坏性的自然灾害，极大地威胁着人们的生命及财产安全。为了应对地震预测这一挑战，北京大学深圳地震监测预测技术研究中心建立了多分量地震监测系统 AETA(Acoustic and electromagnetic testing all in one system)，实时监测电磁扰动和地声信号，并研发预测模型。本文从 AETA 系统的观测数据出发，利用统计学和机器学习相关理论方法，进行临震特征提取，并建立临震预测模型。本文完成的主要工作和创新点如下：

(1) 设计并实现了一种基于主成分分析(Principal Component Analysis)的电磁扰动异常提取算法。提出方法是基于 27 天的太阳周期的时间窗进行主成分提取和矩阵重构，得到背景参考值，然后基于背景参考值和实际值的差值得到异常值。使用该方法在冕宁防震减灾局、九寨沟防震减灾局等台站的 SRSS 波中提取出明显的异常条带，并且该异常条带具备良好的映震效果。

(2) 设计并实现了一种基于 Baer 算子的地声异常提取算法。方法为基于能量及时间域的能量变化使用 Baer 算子构造新的特征序列，而后使用 IQR 方法去除特征序列中的底噪，得到最终异常值。该算法提取的异常值在龙门山断裂带实验中表现出良好的映震效果。

(3) 基于时域、能量、频率特点，选用了 76 种通用统计特征，从多个维度对 AETA 数据进行描述。一方面，这些特征为下游的特征筛选工作提供丰富的候选项;另一方面，由于构建的特征具备实际物理意义，因而特征重要性排序对后续工作具备一定的参考意义。

(4) 提出并实现了一种临震预测模型，对距台站 200km 范围 5 日内的破坏性地震(震级≥Ms5.0)进行预测。在回溯实验中，本文以 2018 年 7 月 1 日至 2018 年 12 月 31 日，地理位置位于川滇地区及其附近的 5 次破坏性地震及距震源 200km 内的 AETA 数据为基础，构建样本集。实验结果表明，构建的临震预测模型查准率可达 0.66，查全率可达 0.75，具备一定的临震预测能力。

综上所述，本文基于 AETA 系统数据提出了临震特征提取及地震预测模型的建立方法，对地震预测进行了有益的探索，并提供了一些可能的研究思路。

关键词：地震预测，AETA，机器学习，主成分分析，特征提取

Research on Feature Extraction and Prediction Model of Imminent Seismic Based on AETA System

Yaxuan Lyu(Microelectronics and Solid-State Electronics)
Directed by Xin'an Wang

ABSTRACT

Megaseism is a highly destructive natural disaster that threatens people's lives and property. In order to meet the challenge of earthquake prediction, Earthquake Monitoring and Prediction Technology Research Center of Peking University Shenzhen Graduate School has established a multi-component seismic monitoring system, AETA(Acoustic and electromagnetic testing all in one system), to monitor electromagnetic disturbances and geoacoustic signals in real time, and to develop prediction models. Based on the observation data of AETA system, this paper extracts the imminent seismic features by using statistical and machine learning theory methods, and establishes an imminent earthquake prediction model:

(1) An algorithm for extracting electromagnetic disturbance based on Principal Component Analysis (PCA) is designed. The proposed method is to extract the principal component and reconstruct the matrix based on the 27-day solar cycle time window, and get the background reference value. Then, based on the difference between the background reference value and the actual value, the abnormal value can be obtained. Using this method, obvious abnormal bands are extracted from SRSS (Sunrise-Sunset) waves of Mianning Earthquake Prevention and Disaster Mitigation Bureau and Jiuzhaigou Earthquake Prevention and Disaster Mitigation Bureau, and the anomalous bands have good seismic reflection effect.

(2) An algorithm for extracting geoacoustic anomalies based on Baer operator is designed. The proposed method is to construct a new feature sequence based on the energy and the energy change in the time domain, and use the IQR method to remove the bottom noise in the feature sequence to get the final outliers. The abnormal values extracted by this algorithm show good seismic reflection in Longmenshan fault zone experiment.

(3) Based on the characteristics of time domain, energy and frequency, 76 statistical features are selected, to describe the AETA data from multiple dimensions, which provides

rich candidates for feature selecting. At the same time, the built features have practical physical meaning. The order of importance of features has certain reference significance for follow-up work.

(4) An imminent earthquake prediction model based on classification model is proposed to predict the destructive earthquakes (magnitude≥ Ms5.0) within 5 days of 200 km from the station. In the retrospective experiment, based on the AETA data within 200 km of five destructive earthquakes in Sichuan-Yunnan region from July 1, 2018 to December 31, 2018, this paper constructs a sample set. The experimental results show that the precision and recall of the imminent earthquake prediction model can reach 0.66 and 0.75 respectively, which means the model proposed in the paper has certain predictive ability for imminent earthquakes.

In summary, based on the data of AETA system, this paper proposes the method of seismic feature extraction and seismic prediction model, which makes a useful exploration of earthquake prediction and provides some possible research ideas.

KEY WORDS: Earthquake prediction, AETA, Machine learning, Principal component analysis, Feature extraction

目录

第一章 绪论 1
1.1 背景及意义 1
1.2 国内外研究现状 2
1.3 本文研究工作 4
1.4 论文的结构 4
第二章 多分量地震监测系统 AETA 及数据预处理 5
2.1 AETA 系统 5
2.1.1 AETA 系统简介 5
2.1.2 AETA 系统电磁扰动信号 6
2.1.3 AETA 系统地声信号 7
2.2 数据预处理 7
2.2.1 缺失值处理 7
2.2.2 异常值处理 8
2.3 本章小结 9
第三章 基于 AETA 数据的临震特征提取 10
3.1 基于地声信号的临震特征提取 10
3.1.1 地声信号的临震异常分析 10
3.1.2 AETA-Baer 异常值计算算法 11
3.1.3 AETA-Baer 方法映震分析 13
3.2 基于电磁信号的临震特征提取 17
3.2.1 电磁信号的临震异常分析 17
3.2.2 滑动 PCA 方法 19
3.2.3 滑动 PCA 方法映震分析 20
3.2.4 PCAETA 的临震特征提取 22
3.3 通用统计特征 23
3.4 本章小结 25
第四章 基于分类模型的临震预测模型研究 26
4.1 样本集的建立 26
4.1.1 地震预测三要素的选取 26

4.1.2 样本不均衡问题及其解决办法 27
4.2 特征选择 29
4.2.1 特征选择方法 29
4.2.2 特征选择实验及结果 30
4.3 分类器的选择 30
4.3.1 决策树 31
4.3.2 支持向量机 34
4.3.3 其它对比算法 38
4.4 模型评估方法及参数优化 39
4.4.1 性能评估方法 39
4.4.2 性能评估指标 40
4.4.3 超参数调优 43
4.5 模型的反演实验 44
4.5.1 样本集构建 44
4.5.2 反演效果 45
4.6 本章小结 45
第五章 总结和展望 46
5.1 总结 46
5.2 展望 47
参考文献 48
攻读硕士学位期间的科研成果 53
致谢 54
北京大学学位论文原创性声明和使用授权说明 57

第一章　绪论

1.1　背景及意义

大地震是一种具有高度破坏性的自然灾害，造成巨大的人员伤亡和财产损失[1]。2009 年 4 月 6 日意大利阿鲁佐（Abruzzo）地区拉奎拉（L’Aquila）发生 Mw6.3 地震，包括昂纳、帕加尼卡和新城堡的历史中心等文化遗址遭到严重破坏或被摧毁。这次强震致使 309 人丧生，1500 人受伤，65579 人被转移，约 20000 幢建筑被破坏或无法居住，经济损失约 30 亿欧元[2]。根据美国地质勘探局(USGS)的官网数据统计，21 世纪（2001-2010 年），全球共发生 232 起存在人员伤亡的地震，共致使 697404 人丧生；伤亡人数在万人以上的地震，平均约 2 年/次，远高于 20 世纪的 5 年/次，地震巨灾呈现频发趋势[3]。如何做好地震预测工作，已成为人民安居乐业的重要课题。

地震预测是固体地球物理学的一个重要课题，通常定义为确定性地表述未来地震的时间、地点和震级[4]。准确地进行地震预测，可以极大地帮助政府决策，及时疏散人群及进行财务转移，最大程度地避免财产损失及人员伤亡。而在长、中、短期和临震预报中，临震预报工作最为重要：一方面，临震预报可以最大程度地减少人员伤亡；另一方面，临震预报即使存在偶尔的虚报现象，也可及时解除，与长、中、短期预报相比，所造成的社会损失要小许多。而作为临震预报的重要实现手段之一，围绕临震特征提取与预测模型的分析与研究具有重要意义。

在 20 世纪 80 年代，科学研究集中在经验分析上，即试图找出地震的独特前兆，或地震前可能发生的一些地球物理趋势或地震活动模式[5]。其基本思想是观察一个前兆，并以高可靠性和准确性发出警报。2011 年，国际民防地震预报委员会（ICEF）审查并确定了一些前兆方法，包括近地表及其上方的电磁变化、地震波速和电导率的变化、地震活动性图像等[6]。近年，国际上杰出的研究成果一般产出于人工智能，即地震前兆与人工智能技术相结合的跨学科研究中。

为了应对地震预测这一挑战，北京大学深圳地震监测预测技术研究中心（下文简称“研究中心”）建立一个广泛的地震数据监测网络——AETA[7]，以作为收集前兆信号的重要手段，和开发预测模型的观测基础。

但是目前，机器学习模型在地震预测方面的成果主要聚焦在在长、中、短期预测，而在临震预测中鲜有报道。而作为一种在各领域表现突出的统计学方法，机器学习在该方向的也许会得到不错的结果。因此，本文试图构建 78 种基于 AETA 数据的临震特征，并使用机器学习方法建模进行临震预测，以期取得一定的效果，并为后来人提供一

些借鉴。

1.2 国内外研究现状

地震预测是一个具有挑战性的话题，全球的研究人员在此方向付出了巨大的汗水与努力[8]。地震预测研究大体可分为三个方面：a)基于历史震例的数理统计 b)前兆信号研究 c)机器学习模型[9]。

基于历史震例的数理统计是指，从历史地震目录中提取指标，并总结该指标序列的规律，从而对未来地震进行预测。该指标一般包括地震频次与震级关系的古登堡-李希特定律（G-R Law）的 a 值和 b 值[10]，地震事件的平均间隔时间、地震能量的平方根变化率、n 次地震事件的平均值震级的均值等。

一些学者在这方面进行了深入研究，如郭等[11]利用本地强震复发的时间间隔，拟合对数正态分布函数，计算未来 50 年的发震概率，得到鲜水河断裂带炉霍段和道孚段未来 50 年大地震的发生概率分别为 0.15 和 0.31；Morales-Esteban 等[12]使用五次地震震级的平均值，第一次地震到第五次地震跨越的时间，以及 b 值变化，使用聚类的方法，尝试对 Ms 4.5 及以上的地震事件进行预测，其敏感性及特异性约为 80~90%。

前兆信号研究即指，试图找出地震的独特前兆，或地震前可能发生的一些地球物理趋势或地震活动模式[5]。其基本思想是观察一个前兆，并以高可靠性和准确性发出警报。1982 年，Raleigh 等人认为，在建立更好的理论模型之前，依赖经验确定的前兆现象仍然是必要的[13]。

21 世纪初，我国的临震预报仍以地震前兆观测为主，前兆信号主体包括：虎皮鹦鹉跳跃现象、次声波、引潮力共振、地应力等信号，若以上主体发生异常，则认为在未来一定范围的时间、地点、震级内会发生地震，在当时 16 次的内部预测试验中,这种方式的成功率可以达到 43.7%，在短临地震被众多学者认为“不可预测”的形势下，达到这样的水平已属不易[14]。2011 年，国际民防地震预报委员会（ICEF）审查并确定了一些前兆方法，包括近地表及其上方的电磁变化、地震波速和电导率的变化、地震活动性图像等[6]。

地震前兆的异常提取通常采用“2s”方法，即观测资料相对正常值的变化超过两倍标准差；或使用相邻观测值之差，即一阶差分，作为临震判别的依据[15]。

随着科学研究、生产消费等社会各领域积累的数据以史无前例的速度迅速膨胀，人们意识到大数据时代已经来临。充分利用海量数据信息，挖掘其中蕴含的价值，已成为学术界、工业界以及各国政府的普遍共识。作为目前主流的信息处理技术，机器学习是实现上述目标的新途径[16]。近年来，机器学习在理论、方法、应用方面取得了卓越的成果。正如 Science 近期发表的综述文章所言[17]，机器学习是当前发展最迅速的信息

科学技术领域之一。目前机器学习已成为求解许多应用问题的主要技术手段，近年，国际上地震领域杰出的研究成果也一般产出于此，即地震前兆与人工智能技术相结合的跨学科研究之中。

目前，机器学习模型在临震预测方面的成果鲜有报道，但它在长、中期地震预测中皆取得了不错的效果，并在短期地震预测中有所进展。

2007 年，Panakkat 和 Adeli[18]提出了一种使用地震活动指标进行地震预测的重要方法[22]。这些指标根据地震的时间分布进行计算，代表该地区的潜在地震状态。作者将得到的 8 个地震参数分别与递归神经网络（RNN），反向传播神经网络（BPNN）和径向基函数（RBF）结合使用，并将该模型应用到南加州及旧金山湾区的预测研究中。结果表明，与其他两个神经网络相比，RNN 具有更好的效果。在此研究之后，2009 年，Adeli 和 Panakkat 将模型更换为概率神经网络(PNN)，使用相同的地震特征对同一区域的地震进行预测，结果显示， PNN 对于震级小于 6.0 的地震具有更好的表现。

2012 年，聂红林等[19]提出了以线性回归及 BP 神经网络技术为基础，对 1970 年 1 月 1 日至 2000 年 12 月 31 日期间，华北、华南两大地震区的交接部位，未来 6 个月地震的震级进行预测。经反演实验，误差分别为±0.78 级及±0.61 级。

2016 年，朱海宁[20]使用改进的支持向量机对我国的最大震级进行预测。该方法以单位时间窗口内的地震频次 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，向后平滑构建数据集。该数据集被喂入 SVM 算法中做回归分析，得到未来一段时间内的最大震级。文章引入了小波变换确定时间窗口长度，方法为对 M-T（最大震级-年份）序列进行连续 morlet 小波变换，再根据小波方差图确定时间序列的活跃周期，即时间窗长。作者将 1900-2000 年的数据作为训练集，2000-2010 年的数据作为测试集。最终报准率达到 93.75%。

2017 年，Asim 等[21]基于 Gutenberg-Richter 反演定律，地震事件的发生频率，前震频率，地震震级分布等地球物理事实，使用数学模型构建了 8 项特征，并输入模式识别神经网络，递归神经网络，随机森林和 LPBoost 中，对 Hindukush 地区未来震级进行预测。次年，Asim 等[22]再次构建了 Gutenberg-Richter 等 60 项地震活动性指标来训练 SVM-HNN 网络，并将该模型应用在 Hindukush、智利和南加州区域，对 M5.0 及以上的地震进行预测。相对于先前的研究，该模型在 MCC，R 评分、准确度等评估标准上得到明显的改善。

总体来讲，国内外关于地震预测的研究中前兆信号的研究成果比较多，取得了可喜的进展，但在地震预测模型与人工智能技术相结合的跨学科研究中，多为长期、中期预测，短临震预测仍处于探索阶段，且在实际的地震短临预测方面，整体水平较低，准确度不高[23,24]，距离真正解决地震三要素难题仍有很大的距离。

1.3 本文研究工作

本文的目的是基于多分量地震监测系统 AETA 的电磁扰动数据和地声数据，进行临震特征提取，并搭建临震预测模型。根据国内外地震预测方法的相关研究，本文主要做了以下几个方面的研究：

(1) 根据 AETA 数据的特点进行数据预处理。本文针对 AETA 电磁扰动数据和地声数据的特点，对断电、断网导致的缺失、突跳数据，分别使用了临近数据和线性插值的方法进行补全，为后续的数据挖掘工作奠定了基础。

(2) 基于 AETA 数据进行特征提取。本文设计了基于主成分分析方法的 PCAETA 特征，和基于 Baer 算子的 AETA-Bear 特征两种人工经验特征，并基于时域、能量、频率特点，选用 76 种通用统计特征，从多个维度对 AETA 数据进行描述，为下游的特征筛选工作提供了丰富的候选项。

(3) 对原始特征进行特征选择。本文使用 Gini 系数、SLR 等特征选择方法对上游提取的 78 个原始特征进行分析，最终选定贡献度最大的 25 个特征，进入下游的分类器进行临震预测。

(4) 搭建临震预测模型。本文对比了决策树、支持向量机、随机森林等分类器，最终选用决策树和支持向量机搭建临震预测模型，对距台站 200km 范围 5 日内的破坏性地震（震级≥Ms5.0）进行预测。在回溯实验中，本文构建的临震预测模型查准率可达 0.66，查全率可达 0.75。

1.4 论文的结构

第一章为绪论，主要介绍了地震预测临震特征提取及预测模型研究的背景与意义，以及国内外的研究情况，并介绍了本文主要的研究内容。

第二章分为两个部分，第一部分对 AETA 多分量地震监测系统进行了系统的介绍，特别是对电磁扰动信号和地声信号来源进行了较为详细的阐述；第二部分根据 AETA 数据的特征进行数据预处理，包括缺失值及异常值的处理。

第三章基于 PCA 及 Baer 算子，提出 2 种基于 AETA 电磁和地声数据的前兆观测方法，并各自提取可用于下游模型的临震特征；另外构建 76 种具有物理意义的统计特征，为特征选择提供丰富的候选项。

第四章介绍了基于分类算法的临震预测模型研究。包括样本集的构建，特征筛选，效果评估及反演实验。

第五章为总结和展望，对本文的研究成果进行了系统总结，并对基于 AETA 数据进行地震临震预测的关键技术进行了展望。

第二章　多分量地震监测系统 AETA 及数据预处理

2010 年起，北京大学深圳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(Acoustic and electromagnetic testing all in one system)，旨在通过低成本、大区域、高密度的设备布设，建立完善的检测网络，捕捉比较一致的前兆异常信号，服务于国内及国际的地震监测预测工作[7]。

2.1　AETA 系统

2.1.1　AETA 系统简介

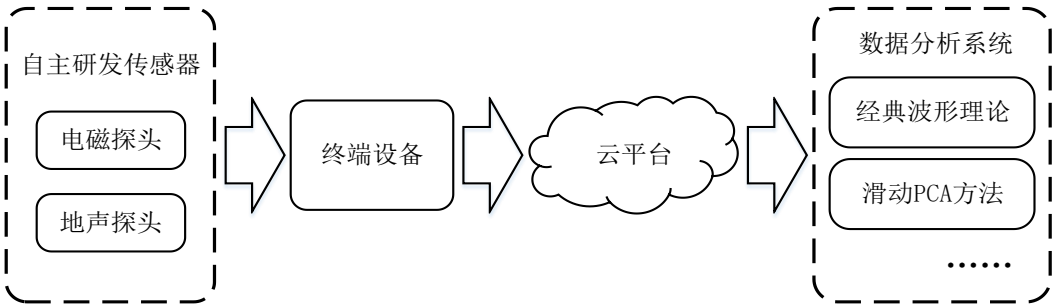


图 2.1　多分量地震监测系统 AETA 系统框图

多分量地震监测系统 AETA 由数据采集系统及数据处理系统组成，数据采集系统包括地声传感探头、电磁传感探头、数据处理终端；数据处理系统包括云平台及数据分析系统，数据流情况如图 2.1 所示[25]。该系统能够感知来自地表及地下的电磁及地声信号，实时采集数据通过互联网网络（有线或无线）将数据传输到云端进行存储及初步特征提取，以便后续进行模型搭建及映震分析。

从全国各台站获取的监测数据，通过互联网络存储在云服务器端，其数据储存形式如表 2.1 所示：

表 2.1　AETA 云服务器中数据储存形式

类型	描述
原始数据	从探头采集到的数据，单位为伏特(V)
均值	单位时间内信号幅值的平均值，用以表示信号能量
振铃技术	单位时间内的过零计数，用以对信号进行频次统计
峰值频率	单位信号内信号的主频率成分

目前，在中国地震局及各高校的支持下，AETA 系统已在全国范围内布设两百余个监测台，遍及北京、四川、云南、台湾等 13 个省市及地区，其中四川省布设达到 111 台，基本覆盖四川省内地震高发区域[26]，如图 2.2 所示。

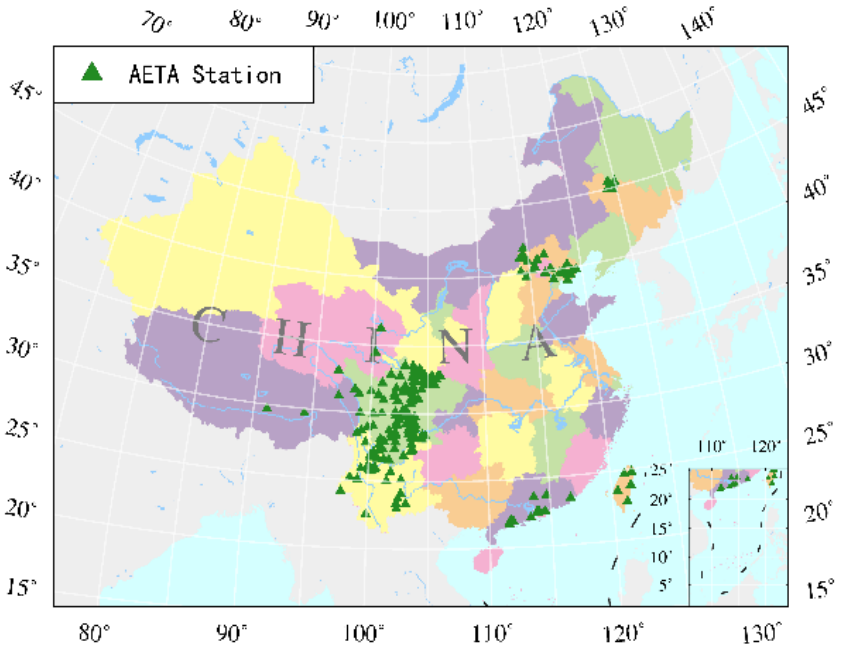


图 2.2　多分量地震监测系统 AETA 布设情况

2.1.2　AETA 系统电磁扰动信号

地震孕育及发生过程中伴随电磁辐射异常已是不争的事实[27-33]，这些异常往往出现在震前数天或数小时[34]。大量观测资料表明，电磁异常是对短临地震反应最敏感的地震前兆之一[33]。

AETA 电磁扰动信号基于法拉第电磁感应定律，当地磁场发生变化时，磁通量相对于感应式磁传感器内固定大小的线圈，做切割磁感线运动，在线圈上产生感应电动势，令闭合的线圈中产生电流。AETA 感应式磁传感器的性能如表 2.2 所示[35]：实物图及测试环境如图 2.3 所示。

表 2.2　AETA 感应式磁传感器的性能

带宽	灵敏度	噪声水平	分辨率	采样率
0.1Hz~10Hz	20mV/nT @0.1Hz~10kHz	0.1~0.2 pT/Hz1/2 @(10 Hz~1 kHz)	18bit	低频 500Hz 全频 30kHz

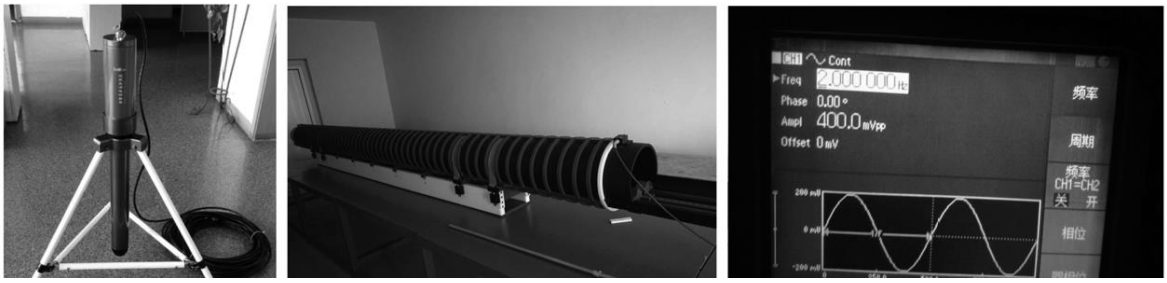


图 2.3 电磁探头实物图以及灵敏度、噪声水平的测试环境

2.1.3 AETA 系统地声信号

地声来源于地下，是地壳运动的直接现象。历史上多次大震前，均有地声异常的记载，如轰隆声、打雷声等。近代的地声观测记录及分析表明，地震发生前几小时及小时常伴有地声在时域及频域的异常变化。AETA 地声探头[36]性能如表 2.3 所示：

表 2.3　AETA 地声探头的性能

带宽	灵敏度	工作温度	分辨率	采样率
0.1Hz~50kHz	3LSB/pa @0.1Hz~50kHz	-50~100℃	18bit	低频 500Hz 全频 150kHz

2.2 数据预处理

由于不可抗力的因素，停电、断网等等，造成了数据的缺失和突跳，将对数据挖掘的质量及结果的稳健性造成影响，因此对数据进行预处理是十分必要的。

2.2.1 缺失值处理

缺失值处理主要有几种方法[37-38]：

(1) 删除法，即将数据表中含有缺失值的数据删除；

(2) 插补法，即基于某种规则用该空最有可能的值进行填补，该规则根据数据类型及分布，可使用：

① 均值插补。若缺失类型为数值型数据，可使用平均值填补；若缺失类型为类别性数据，可使用众数进行填补。

② 回归插补。

③ 极大似然估计。极大似然估计是指，根据观测数据的分布推测模型中的未知参数，这里需要缺失值为随机缺失。该方法适用于大样本，以使得观测样本数足以保证极大似然法的估计值逼近无偏且服从正态分布。

④ 多重插补。该方法的实现类似于强化学习，先估计出一个背景值，然后在这

个背景值上加入噪声，构成多个候选项，最后经过事先设定好的规则，对候选项进行选取。

⑤ 其它插补方法。如 RAR 方法、MVC 方法、FRCAR、GBARMVC 等基于关联规则的插补。

(3) 对于规则数据，使用临近数据进行填补。

AETA 电磁数据具有明显的日周期现象，也就是说，在无震情况下，相邻两日对应时间的数据相差不大，因此使用第 3 种方法进行填补，填补效果如图 2.4 所示，蓝色为原始数据，橙色为填补的数据。

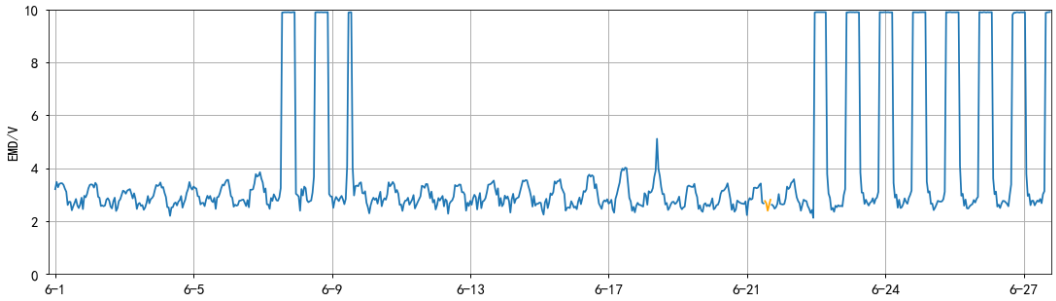


图 2.4　电磁数据填充样例-冕宁防震减灾局台站

AETA 地声数据的特点为一般情况下，信号幅值较低，变化缓慢，因此地声数据使用曲线回归进行填补，填补效果如图 2.5 所示，蓝色为原始数据，橙色为填补的数据。

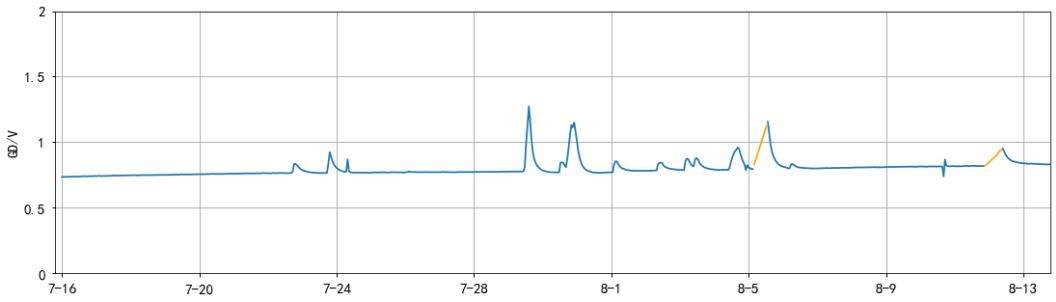


图 2.5　地声数据填充样例-青川防震减灾局台站

2.2.2　异常值处理

在实际场景中，台站会出现断电的情况，在电力恢复后，AETA 地声信号受电容影响，在一段时间内会出现明显且规律的异常，表现为均值信号会在短时间内从高位向下回落，如图 2.6 所示。

对此类异常，本文的解决方法为，以前 27 日的最高幅值为阈值，删去异常信号中高于阈值的部分。

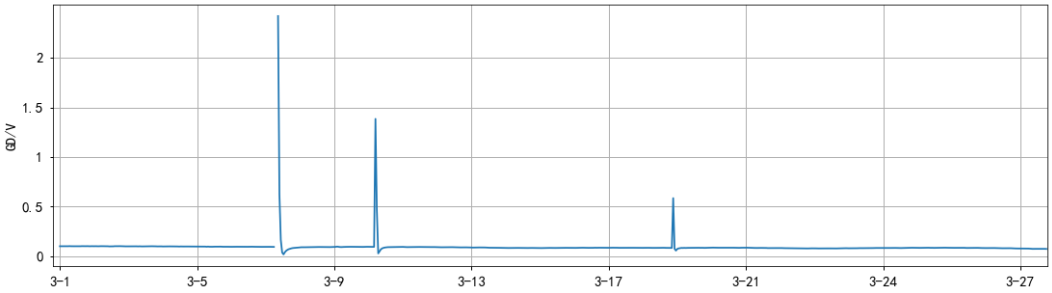


图 2.6　地声数据异常样例-名山县安吉村台站

2.3　本章小结

本章分为两个部分，第一部分对 AETA 多分量地震监测系统进行了系统的介绍，特别是对电磁扰动信号和地声信号来源进行了较为详细的阐述；第二部分根据 AETA 数据的特征进行数据预处理，包括缺失值及异常值的处理，该操作有助于令下游任务的分析结果更加准确。

第三章　基于 AETA 数据的临震特征提取

在特征提取的过程中，本文考虑了在实践过程中，对临震信号进行观察获取的经验总结，所获取的人工特征，也充分利用了前人总结的通用特征所带来的便利性，尽可能为下游模型提供丰富的候选项。

如第二章所述，AETA 系统的原始信号基于两种物理源，电磁扰动信号和地声信号，在实际观测中电磁信号不断波动，地声信号一般情况下比较安静，偶尔波动。

研究中心认为电磁信号出现 SRSS 波后，地震开始孕育[41]，因而电磁信息中应当隐藏着孕震的能量信息，本文使用滑动 PCA 方法对 SRSS 波进行处理，可在热力图中观察到震前的条带异常现象，该方法可作为一种地震前兆预测方法；同时根据条带异常的映震关系，生成了基于 PCAETA 的临震特征。

同时，对于地声信号，本文也进行了处理。在预测实践中发现，地声信号往往在震前呈现幅度及频率增大的现象，因而本文借用 Baer 在改进 STA/LTA 方法时使用的算子，对地声数据进行处理，结果发现相对于单纯使用幅度进行计算，Baer 特征具有更好的映震关系，可作为后续模型的临震特征之一。

此外，本文生成 76 种常用的时间序列统计特征，对电磁及地声信号进行特征提取，为后续特征选择提供了丰富的候选项；并且，这些统计特征具有明确的物理意义，其值的变化对于日后地球物理工作的展开，具有一定的参考作用。

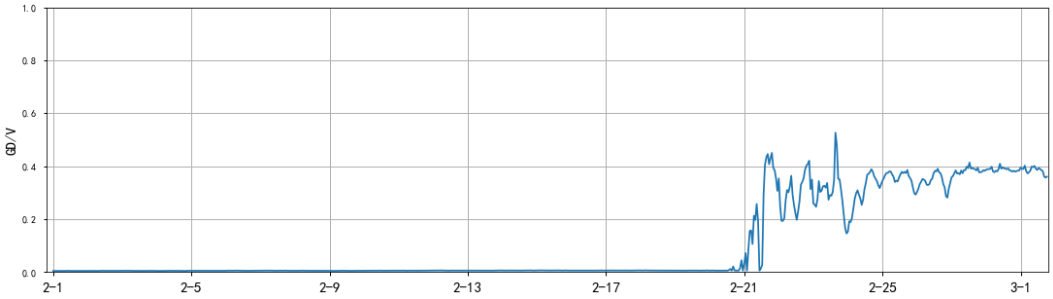
3.1　基于地声信号的临震特征提取

本节分为三个部分，第一部分给出地声信号临震异常的案例分析，从而总结出地声信号临地震异常的特点；第二部分给出临震异常信号的提取算法，该算法得到的临震异常可作为地震预测的参考依据，同时该异常值也可作为特征输入到下游模型中；第三部分为实验部分，验证该算法的映震效果。

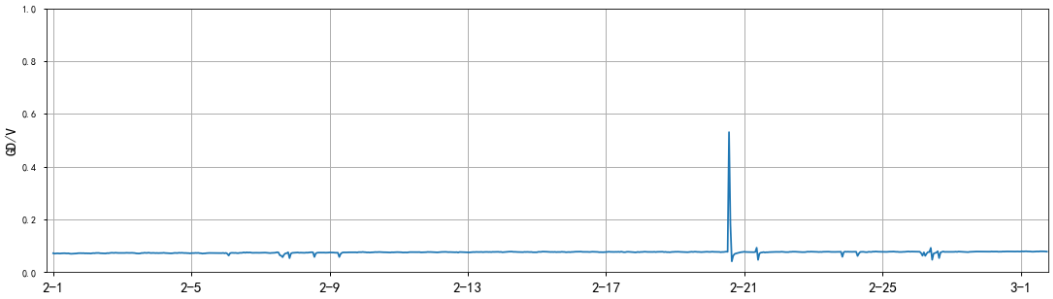
3.1.1　地声信号的临震异常分析

图 3.1 给出 3 个地震与台站对应关系的案例分析。2019 年 2 月 23 日，北京时间 1 时 42 分 58 秒四川广元市青川县发生 3 级地震。图 3.1(a)及图 3.1(b)给出该地震附近 2 个台站，AETA 低频地声均值信号的变化情况。可以看到，在地震发生前两日，广元朝天区东溪河台信号出现强烈的扰动，表现为幅度明显抬高，且振动频率增大；平武白马乡的信号则出现单脉冲信号。图 3.1(c)给出 2017 年 7 月 8 日至 8 月 7 日的位于青川防震减灾局的 AETA 低频地声均值信号。在该时段距台站 200km 范围内，共发生 3 次地

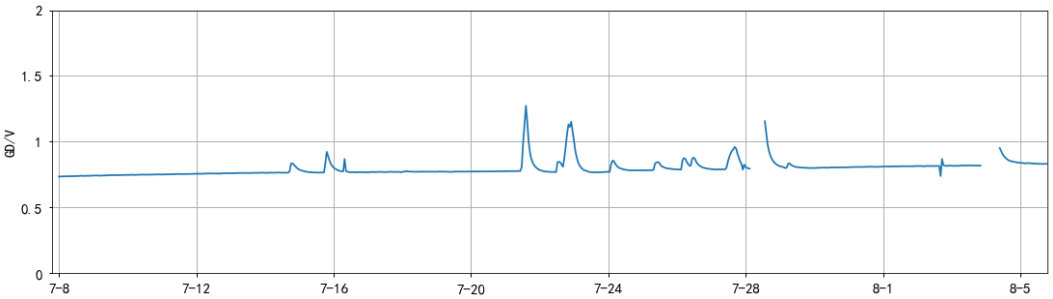
震，分别为 2017 年 7 月 17 日青川县的 4.9 级地震，7 月 20 日再次的 3 级地震，7 月 26 日附近平武县的 2.9 级地震。这三次地震地震前后均出现信号扰动的现象。



(a)　2019-02-01 至 2019-03-03 广元朝天区东溪河台低频地声信号



(b)　2019-02-01 至 2019-03-03 平武白马乡低频地声信号



(c)　2017-07-08 至 2017-08-07 青川防震减灾局低频地声信号

图 3.1　低频地声均值信号临震案例分析

3.1.2　AETA-Baer 异常值计算算法

根据历史震例，可以发现 AETA 地声信号的临震异常往往同时表现出强度和频率的增强，于是这些特征可以拟合出更为明显的特征。另外，由于地面声音所在频段存在各种噪声源，如雨声、车辆声、爆破声、有线广播、无线广播等，会影响地面声音的记录和产生噪声。因此，本节提出的异常检测方法可分为两部分：计算新的特征序列及去除底噪，获取异常指标。

3.1.2.1　计算新的特征序列

Baer 和 Kradolfer[39]结合微震信号的特点，利用能量和时间域的能量变化构造特征序列，将波形特征序列应用于长短时窗法，形成一种更有效的 STA/LTA 方法。该方法已成为地震信号自动识别领域中广泛应用的经典算法之一。微地震记录的特征是高频率、低信噪比，这与地声前兆数据的特征是一致的。因此，本文尝试使用相同的特征函数来处理地声数据。特征序列定义为

$$CF_{(i)Baer} = y_{(j)}{}^2 + K^2\left[y_{(j)} - y_{(j-1)}\right]^2 \tag{3.1}$$

$$K_{(i)Baer} = \frac{\Sigma_{j=1}^{i}|y_{(j)}|}{\Sigma_{j=1}^{i}|y_{(j)}|} = \frac{\Sigma_{j=1}^{i}|y_{(j)}|}{\Sigma_{j=1}^{i}|y_{(j)}-y_{(j-1)}|} \tag{3.2}$$

式中 k 为加权因子，y 为振幅，$y_{(j-1)}$为振幅的一阶差。该特征序列具有振幅和频率两个特征，能很好地反映它们在时间域中的同步变化。

3.1.2.2 得到特征序列的异常值

在统计学中，有多种指标可用于描述数据形态，常见的如平均值、标准差、方差、中位数和四分位数间距等。标准差和方差常用于描述正态分布或近似正态分布的场景，而平均值、方差、中位数及四分位数间距则可用于大多数场景，如偏态分布、未知分布等，其中，中位数及四分位数可以消除极端数据的影响，使得均值和标准差更稳定[40]。由于地声数据的复杂性，其在不同时期的数据分布仍存在不确定性，因此，本文使用四分位数方法来处理数据。

在地震前兆数据分析领域，滑动四分位距算法是一种流行的算法[41-43]。在 AETA 系统中，本文将之应用于低频和全频电磁扰动数据均值的分析上。滑动四分位距法的重要特征是有一个滑动窗口，该滑动窗口大小的设置根据应用场景具体设定。

当分析一日的特征数据时，可以以时间窗口为单位构建一个数据矩阵 $X = \begin{pmatrix} x_{1\times1} & \cdots & x_{1\times n} \\ \vdots & \ddots & \vdots \\ x_{m\times1} & \cdots & x_{m\times n} \end{pmatrix}$，其中每一列为一小时的均值数据，m=24。考虑到地球声音的变化会受到潮汐涨落等地球活动的影响，将滑动窗周期 n 设为 27，即太阳活动周期，可以消除太阳活动的干扰。

将矩阵每行按升序重构，得到三个节点：滑动窗口中的第一个四分位数（25%）记为 Q1，第二个四分位数（50%），记为 Q2，第三个四分位数（75%），记为 Q3，则四分位数间距 IQR 计算如下：

$$IQR = Q3 - Q1 \tag{3.3}$$

四分位距 IQR 大致等于 1.34 倍的标准差，Q2 常常用于预测数据。则可得异常的上限 UB(UP Boundary)和下限 LB（Low Boundary）得到定义，如下式：

$$up = Q2 + k * IQR \tag{3.4}$$

$$low = Q2 - k * IQR \tag{3.5}$$

在大多数情况下，参数 k 设置为 1.5。

$$OUTLIER = \begin{cases} CF - up; & while\ CF > up \\ 0\ ; & while\ low < CF < \text{up} \\ CF - low; & \text{while}\ CF < low \end{cases} \tag{3.6}$$

$OUTLIER$描述了地声扰动的异常程度。如果实际 Baer 特征不超过上下限，则$OUTLIER$=0，表示没有异常。如果实际 Baer 特征高于上限，$OUTLIER > 0$，表示为正异常。如果实际 Baer 特征低于下限，则$OUTLIER < 0$，表示为负异常。

3.1.3　AETA-Baer 方法映震分析

3.1.3.1　实验区域

龙门山断裂带是青藏高原与川西前陆盆地之间的边界断裂，控制着龙门山地区的地震活动。龙门山长约 500 公里，宽约 30～40 公里[44]。近 30 年来，许多学者对龙门山地区的应力应变状态和电磁干扰的变化进行了研究，结果表明，龙门山地区这些能量场的变化具有一定的映震效果[45-47]。

为了保证数据质量，本文选择了龙门山北站作为实验区。数据来源于该区域的四个站点，其地理位置如表 3.1 所示。

表 3.1　实验区内各监测点及其基本情况

ID	监测点	缩写	地理位置	震中距(km)
90	茂县测点	MX	33.25°N, 104.24°E	40
198	平武白马藏族乡	PW	32.65°N, 103.60°E	64
150	青川房石	QF	32.41°N, 104.55°E	111
19	都江堰中学	DJY	32.59°N, 105.23°E	147

3.1.3.2　实验结果及分析

以 DJY 台站为例，图 3.2（a）显示了该站 2017 年 9 月 1 日至 2018 年 2 月 28 日收集的 GD 数据（低频地声均值数据），其中缺失数据为空白。利用 Baer 算子处理图 3.2（a）所示的 GD 数据，可以得到 GD 数据的 Baer 特征的日变化波形，如图 3.2（b）所示。图 3.2（b）中 Baer 特征的异常指标采用滑动四分位数法得到，结果如图 3.2（c）所示，图 3.2 中的黄线表示 2018-05-16 平武地震（Ms4.0）。

利用该方法可以得到 4 个台站同一时期的地声异常指标，结果如图 3.3 所示，为方

便对比异常与地震见的对应关系，图中标注了在此期间台站附近发生的地震，地震标注方法见表 3.2，震级大，线越长。

图 3.3 显示了 2018 年 5 月 1 日至 2018 年 9 月 31 日 4 个台站的地声变化异常指数，以及在此期间台站附近发生的地震。据统计，实验期间，台站附近的地震异常共 26 处，其中发生在地震前后 7 天的为 16 处，占异常总数的 61.54%；台站附近共发生了 13 次地震，其中 11 次地震可在附近找到至少一个异常台站，占地震总数的 84.62%，地震基本情况如表 3.3 所示。实验结果表明，该方法具有一定的映震能力。

表 3.2　地震标注方法

颜色	震级	震中距
红色	≥Ms2.0	≤100km
黄色	≥Ms4.0	100~300km
蓝色	≥Ms8.0	≤1000km

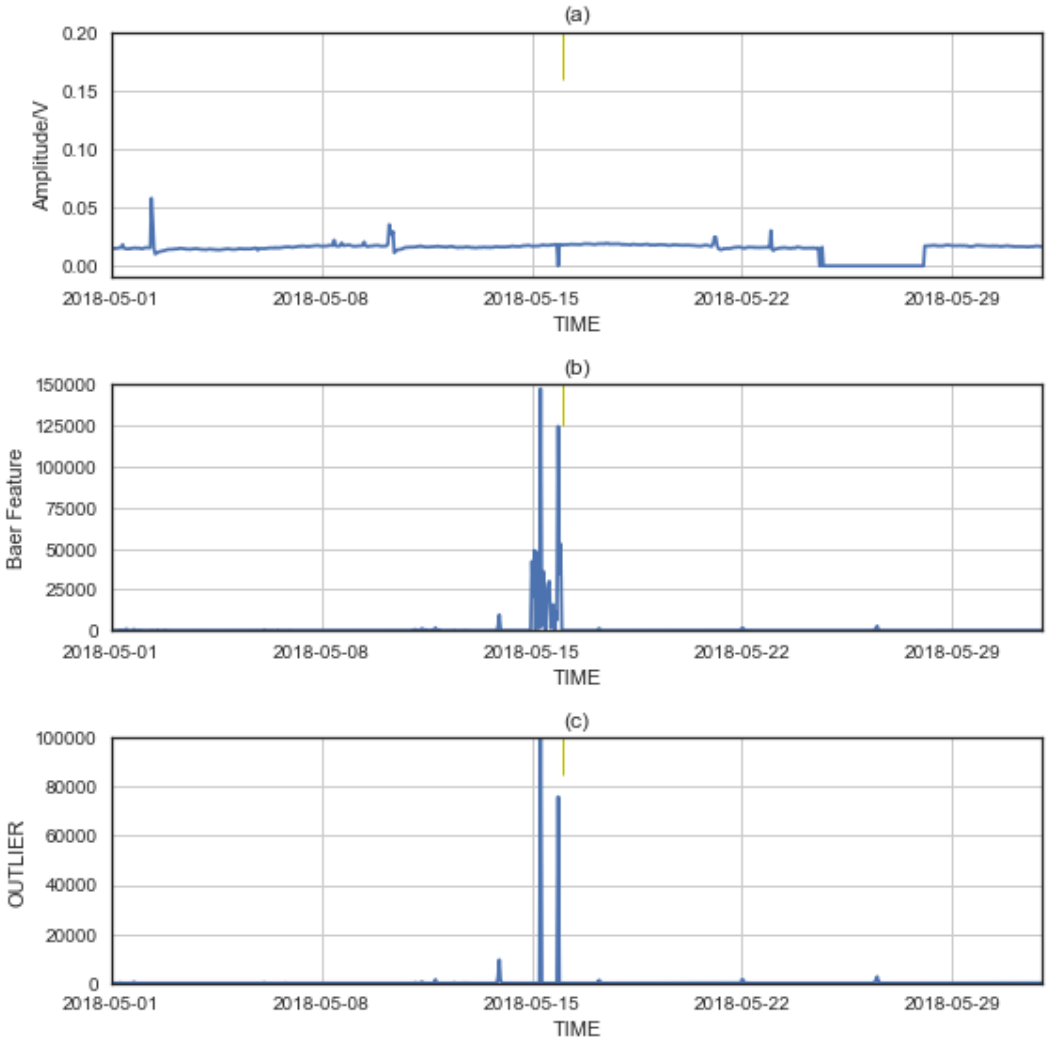


图 3.2　(a)DJY 台站的 GD 数据 (b)DJY 台站的 Baer 特征数据 (c)DJY 台站的异常值

为了探讨 Baer 算子的有效性，本文对相同台站相同时间段的数据，使用 IQR 方法直接处理振幅数据，结果如图 3.4 所示，地震标记方法与图 3.3 相同。

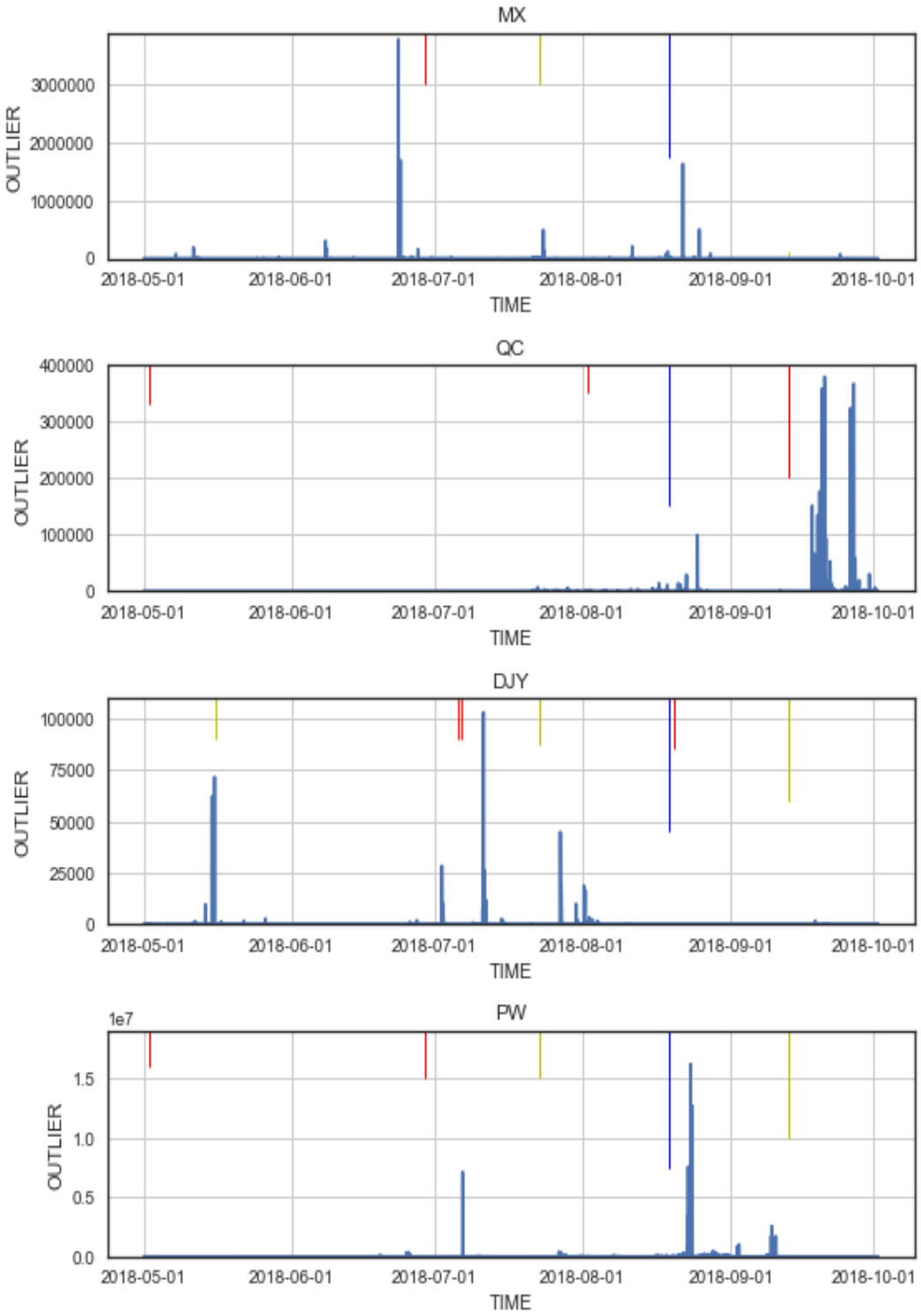


图 3.3　四个台站的 AETA-Baer 异常指标

图 3.4 表明，滑动 IQR 法处理的振幅数据在一定程度上反映了数据的异常情况，特别是 2018-08-02 陇南地震（MS2.9）对应的 QC 台异常，2018-07-06 成都地震（MS3.1 和 MS3.6）对应的 DJY 台异常，2018-06-29 平武地震（MS4.0）和 2018-09-12 宁强地震（MS5.3）对应的 PW 站异常。然而，用这种方法算得的异常时间占比太大，达到 37.89%，尤其是在 MX 台中达到 74.43%；而 Baer 算子算得的异常占比分别为 19.72% 和 27.22%。此外，用 Baer 算子算得的异常，与地震具有更好的对应关系。由此可见，Baer 算子对于 GD 数据具备有效的异常提取能力。

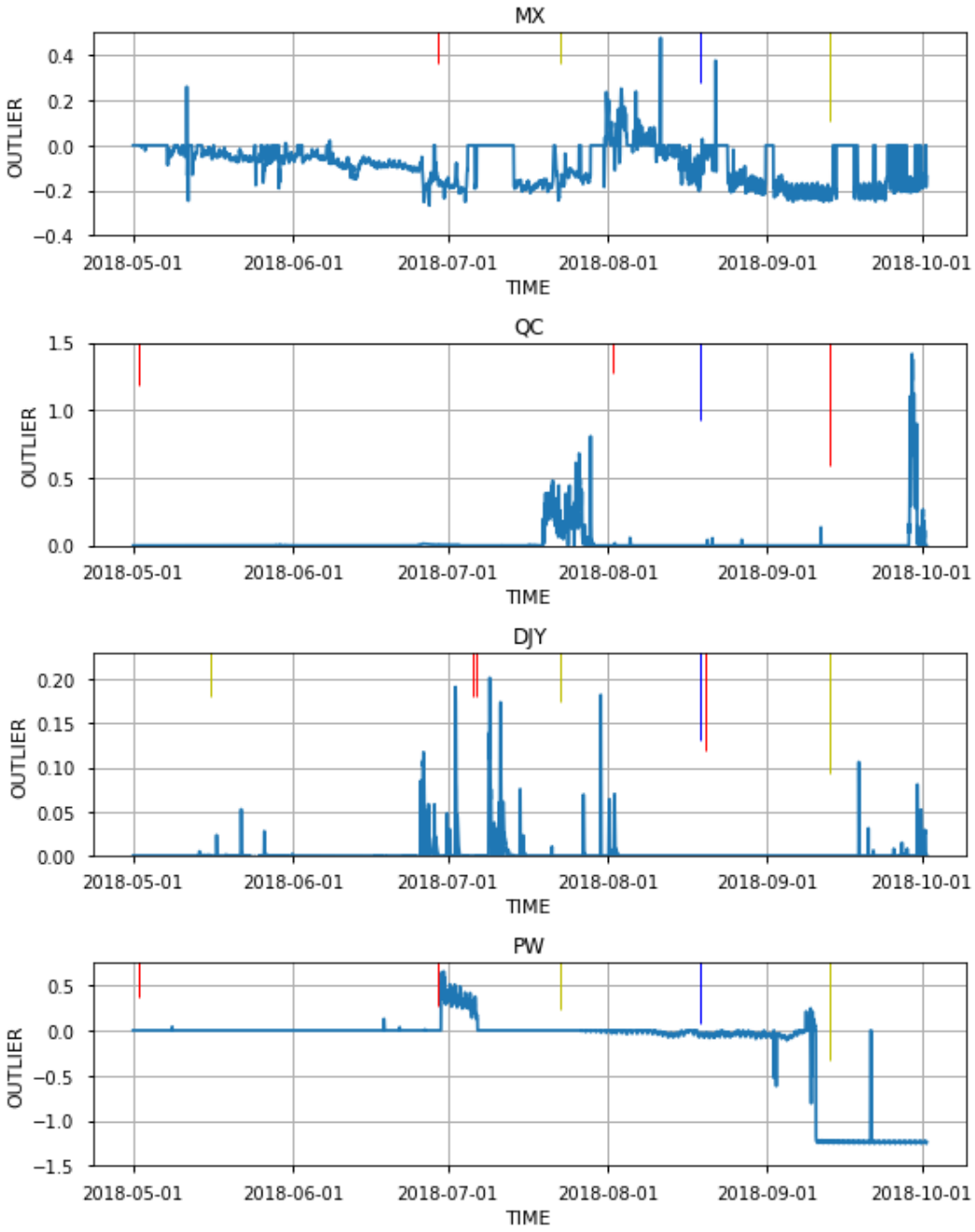


图 3.4　四个台站的滑动 IQR 异常指标

表 3.3　实验区内地震情况

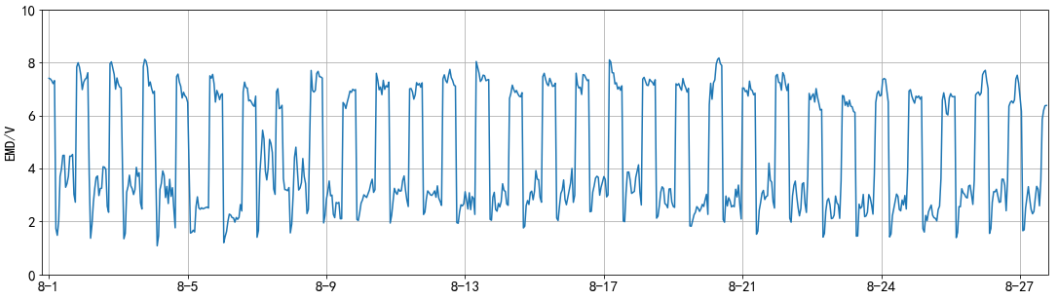
时间	震级	纬度	经度
2018-05-02 07:10:51	M3.3	32.60°N	105.37°E
2018-05-16 16:46:40	M4.3	29.18°N	102.27°E
2018-05-16 16:46:11	M4.3	29.19°N	102.28°E
2018-05-16 16:44:02	M3.2	29.20°N	102.26°E
2018-06-29 08:42:18	M4.0	32.17°N	104.57°E
2018-07-06 12:49:46	M3.1	20.35°N	103.29°E
2018-07-06 13:19:45	M3.6	20.36°N	103.28°E
2018-07-23 07:02:03	M4.2	29.55°N	104.55°E
2018-07-23 06:43:49	M3.6	29.55°N	104.54°E
2018-08-02 22:34:21	M2.9	32.97°N	105.59°E
2018-08-19 20:31:17	M2.8	20.37°N	103.30°E
2018-08-19 08:19:37	M8.1	18.08°S	178.06°E
2018-09-12 19:06:34	M5.3	32.75°N	105.69°E

3.2　基于电磁信号的临震特征提取

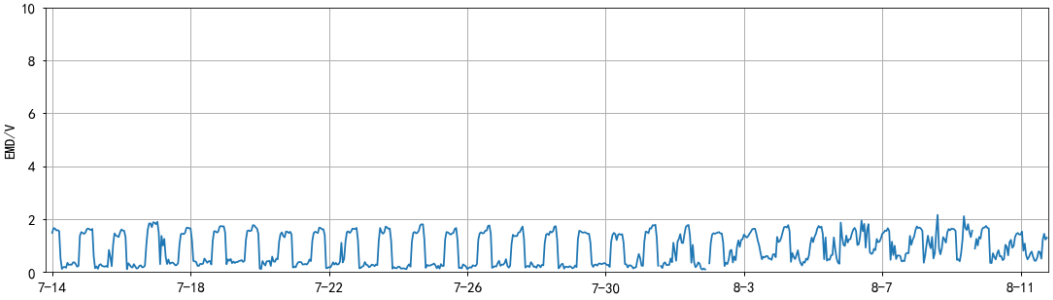
本节分为四个部分，第一部分给出电磁信号临震异常的案例分析，从而总结出电磁信号临地震异常的特点；第二部分给出临震异常信号的提取算法，该算法得到的临震异常可作为地震预测的方法之一；第三部分为实验部分，验证该算法的映震效果；第四部分给出从该算法中定量提取临震特征的方法。

3.2.1　电磁信号的临震异常分析

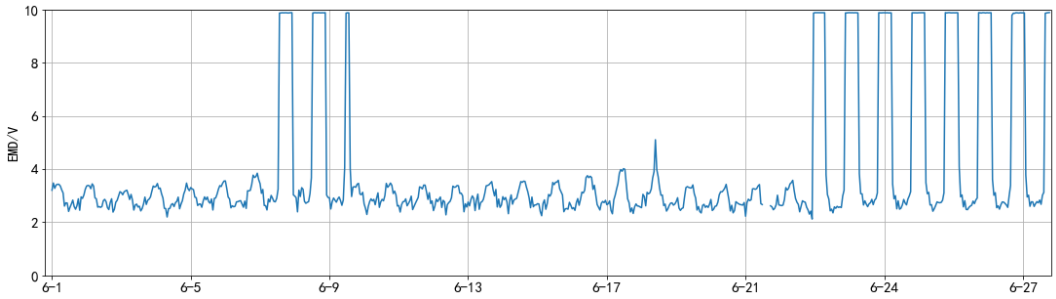
图 3.5 给出 2017 年 2 个地震与台站对应关系的案例分析。2017 年 8 月 8 日，北京时间 21 时 19 分 46 秒，四川省北部阿坝州九寨沟县发生 7.0 级地震。图 3.5(a)和图 3.5(b)给出该地震附近台站，AETA 低频电磁扰动信号的变化情况。可以看出，地震发生时，九寨沟防震减灾局的信号在当月呈现规律的高幅值方波波动，茂县测点也出现日周期的方波信号。图 3.5(c)表示，2017-06-09 四川省宜宾市发生 Ms3.4 级地震，在地震前后一日，冕宁防震减灾局信号持续出现高幅值方波波动。在实际分析工作中，也发现了很多类似的案例，即在地震前后，震源附近台站有较大概率出现方波波动，本文认为该方波是一种具有标志性的信号，预示着台站附近将发生地震。



(a) 2017-08-01 至 2017-08-27 九寨沟防震减灾局低频电磁信号



(b) 2017-07-14 至 2017-08-11 茂县测点低频电磁信号



(c) 2017-06-01 至 2017-06-30 冕宁防震减灾局低频电磁信号

图 3.5　低频电磁均值信号临震案例分析

特别的是，九寨沟防震减灾局的电磁均值出现一种与日升日落几乎同步的波动，日升时变低，日落时变高，研究中心将这种波定义为[48]“SRSS 波”。图 3.6 表示九寨沟县的日升日落时间与波形上升下降沿时间的对比。

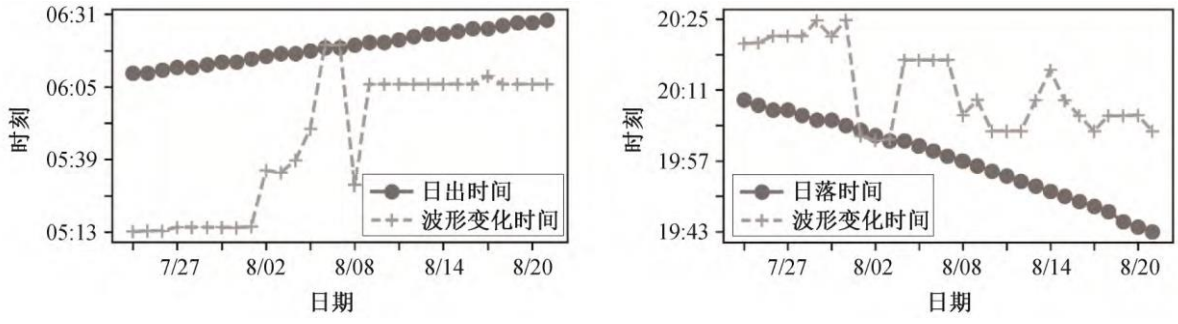


图 3.6　九寨沟 SRSS 波变化时间点与日升日落时间对应关系

3.2.2　滑动 PCA 方法

PCA 应用于震前 TEC 异常监测方面取得一些进展。2016 年，邹斌等[49]将 PCA[50]、滑动时窗法和张小红等提出的限差确定策略[51]相结合，提出一种震前电离层异常探测新方法：滑动 PCA 方法。本文将滑动 PCA 方法运用到 AETA 的 SRSS 波的异常分析中，绘制热力图，并根据临震特征构建 PCAETA 异常值，进行地震预测。

3.2.2.1　参考背景值计算

设时间序列矩阵

$$\mathrm{X}=\begin{pmatrix} x_{1\times1} & \cdots & x_{1\times27} \\ \vdots & \ddots & \vdots \\ x_{24\times1} & \cdots & x_{24\times27} \end{pmatrix} \tag{3.7}$$

表示时长 27 日，每日数据时间间隔为 1 小时的原始数据。

计算得到原始矩阵 X 的协方差矩阵 C，求得协方差矩阵的特征值 λ1≥λ2≥…≥λn 及对应的特征向量矩阵 E,则得到

$$X=PE \tag{3.8}$$

其中 P 可表示为$P=(P_1,P_2,\cdots,P_n)$，E 可表示为$E=(E_1,E_2,\cdots,E_n)$

选取前 k 个特征值，使累计贡献率$\alpha \geq 85\%$，其中α 为

$$\alpha=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\times100\% \tag{3.9}$$

设置$P^*=(P_1,P_2,\cdots,P_k)$，　$E^*=(E_1,E_2,\cdots,E_k)$，计算重构矩阵

$$X_p=P^*E^* \tag{3.10}$$

选取重构矩阵X_p最后一列X_{ref}为当日背景参考值。

3.2.2.2　异常判定

为进一步防止 AETA 系统在运行过程中可能受到偶然的环境干扰的影响，提升 PCA 预测的鲁棒性，需设立合理的误差范围：

计算当日实际值X_{true}与参考值X_{ref}差值的绝对值，按从小到大排列，选取第 23 位（95.8%），得到公差Δ，并以此进行异常判定。

计算异常限值：

$$\begin{cases} up=X_{ref}+\Delta \\ low=X_{ref}-\Delta \end{cases} \tag{3.11}$$

计算异常程度：

$$\Delta X_{abn} = \begin{cases} X_{true} - up, & X_{true} > up \\ 0, & low < X_{true} < up \\ X_{true} - low, & X_{true} < low \end{cases} \tag{3.12}$$

若$\Delta X_{abn} > 0$，认为存在正异常；若$\Delta X_{abn} < 0$，认为存在负异常。

3.2.2.3 滑动 PCA 方法探测流程

首先利用 PCA 方法构建主成分模型，得到目标日期的参考背景值后，探测该日异常程度，然后将样本向后滑动 1 日继续计算，直至计算完毕，最后利用所得到的异常值绘制热力图，观察异常情况。

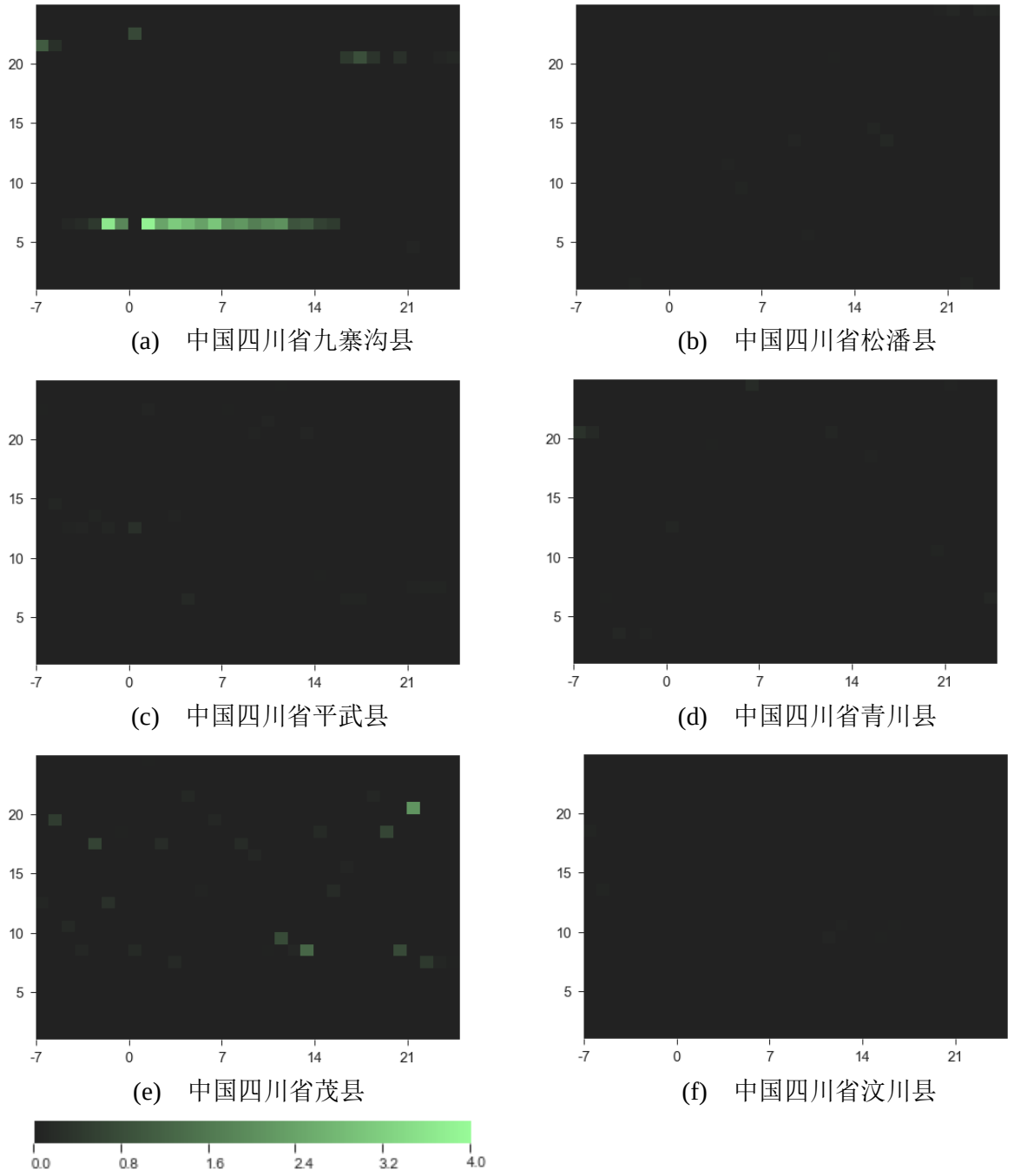
在实际情况中，数据可能由于监测台断电、断网等原因出现短时空缺现象。如果样本数据存在不超过 7 日的短时空缺，则使用相邻日期对应时间数据进行填补，保证数据结构及长度一致即可，不影响预测结果。

3.2.3 滑动 PCA 方法映震分析

表 3.4 表示 2017 年 8 月 8 日九寨沟 Ms 7.0 地震发生时，距离震源 200km 内的 AETA 台站及其基本信息。图 3.7 及图 3.8 表示使用滑动 PCA 算法，对各 AETA 台站中 AETA 设备的观察数据进行计算，得到各台站 2017 年 8 月 1 日至 2017 年 8 月 31 日的异常值热力图，及九寨沟防震减灾局 2017 年 7 月 11 日至 2017 年 10 月 19 日的百日异常值热力图。可以看出，地震发生前 5 日，每天 06-07 时出现异常程度逐渐增强的连续异常条带，地震发生前 2 日，异常点的异常程度大于 2，明显高于历史背景点中其他异常点；地震于 21 时 19 分发生，异常点位于当日 21-22 时，表明此时的异常程度大于该日其它时段；地震发生后，06-07 时再次出现一场程度逐渐减弱的连续异常条带，持续 15 天，也就是余震频发的时期，之后没有大的余震发生。而震源附近其它台站的一场程度小且异常程度散乱，没有构成异常条带。另一方面，九寨沟防震减灾局的条带异常结束后，未形成新的高幅值连续条带，对应时间 200km 内亦无强震。

表 3.4　2017 年九寨沟 Ms7.0 地震 200km 内各监测点及其基本情况

监测点	名称缩写	安装时间	地理位置	震中距(km)
九寨沟	JZG	2017-06-10	33.25°N, 104.24°E	40
松潘	SP	2017-06-12	32.65°N, 103.60°E	64
平武	PW	2017-06-08	32.41°N, 104.55°E	111
青川	QC	2017-06-07	32.59°N, 105.23°E	147
茂县	MX	2017-06-13	31.69°N, 103.85°E	167
汶川	WC	2017-06-13	31.48°N, 103.59°E	192



(a)　中国四川省九寨沟县　　(b)　中国四川省松潘县

(c)　中国四川省平武县　　(d)　中国四川省青川县

(e)　中国四川省茂县　　(f)　中国四川省汶川县

图 3.7　M7.0 九寨沟地震附近监测台异常值热力图(UTC+8)

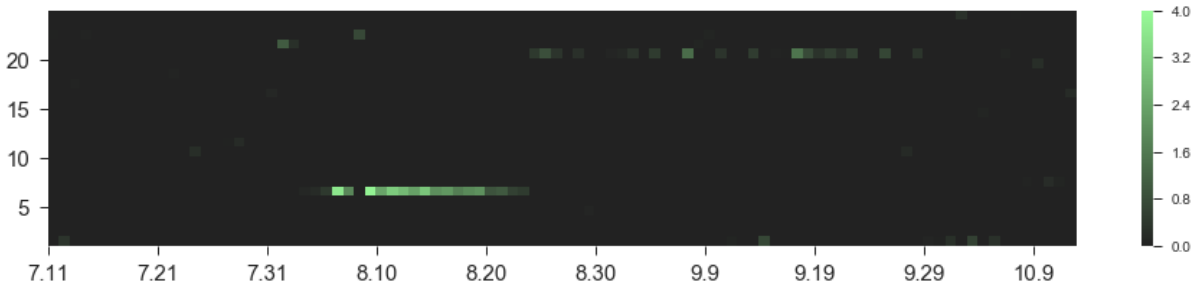


图 3.8　九寨沟 AETA 台站异常值热力图

由 PCAETA 算法得到的临震异常条带及其映震效果,在冕宁防震减灾局 AETA 台站(MN)的 AETA 观测数据中得到进一步验证,图 3.9 表示冕宁防震减灾局 AETA 台站 2017 年 8 月 12 日至 2017 年 11 月 20 日的百日异常值热力图。表 3.5 列出 2017－08－12 至 2017－11－20 距冕宁防震减灾局 AETA 台站 200 km 范围内地震及其与条带异常的对应情况,可以看出,在该时段内,如果条带异常发生在 2~3 天后,则一定会发生地震;如果没有条带异常，则没有大于 3.0 级的地震。

综上所述，经 PCAETA 处理得到的的条带异常现象是比较明确的临震前兆特征。

表 3.5　2017－08－12 至 2017－11－20 距离冕宁防震减灾局 AETA 台站 200km 内地震及其与条带异常对应情况

条带异常时段	发震时刻(UTC+8)	震级	震中位置	深度/km	震中距/km
2017-08-27 至 2017-09-16	2017-08-30 13:46:17	Ms 3.0	27.90°N, 101.37°E	8	107
	2017-09-12 18:40:10	Ms 3.2	27.92°N, 101.42°E	13	101
	2017-09-12 19:25:59	Ms 3.0	27.91°N, 101.39°E	15	101
	2017-09-12 19:26:40	Ms 4.4	27.93°N, 101.42°E	13	101
2017-09-19 至 2017-10-09	2017-09-21 10:03:12	Ms 3.0	27.93°N, 101.41°E	15	101
2017-10-14 至 2017-11-01	2017-10-25 14:29:37	Ms 3.0	27.90°N, 101.43°E	13	102

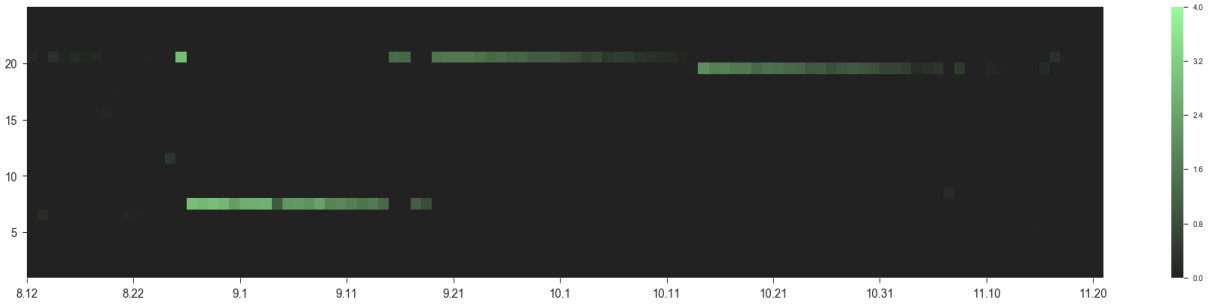


图 3.9　冕宁防震减灾局 AETA 台站异常值热力图

3.2.4　PCAETA 的临震特征提取

3.2.3 小节中所述方法可作为一种临震时间预测依据，但若将该异常值直接输入下游模型，则可能无法学习“异常条带”这一特征。因而，本文对异常条带的持续时长做出统计，并以此作为下游模型的特征。

3.3　通用统计特征

本文选用了 38 种基于时间序列的统计特征提取方法，覆盖信号的频率、功率和熵等特性。其中包含 30 种常规统计特征提取方法，如表 3.6 所示，及 8 种与自回归等性质有关的提取方法。

在生成特征时，本文使用窗长为 27d 的窗口在时间序列上进行滑动，并对窗口内时间子序列进行处理，得到每日的统计特征。

本文将该 38 种统计特征提取方法应用于 AETA 低频电磁均值及低频地声均值数据中，故共计 76 种特征。

表 3.6　常规统计特征

序号	物理意义
1	和
2	平方和
3	绝对值之和
4	均值
5	方差
6	中位数
7	最大值
8	最小值
9	熵
10	峰值个数
11	是否存在重复值
12	大于均值的点的个数
13	小于均值的点的个数
14	大于均值的最长子序列长度
15	小于均值的最长子序列长度
16	最大值首次出现的位置
17	最小值首次出现的位置
18	最大值是否存在重复值
19	最小值是否存在重复值
20	将时间序列中的点分为 4 份(即四个星期)后求熵

续表 3.6　常规统计特征

序号	物理意义
21	出现超过 1 次的值的个数/总的非空值个数
22	出现超过 1 次的值的个数/总个数
23	频繁点的个数
24	频繁值之和
25	方差是否大于标准差
26	基于迪克－富勒检验，检查时间序列中是否存在单位根
27	近似熵，可以衡量一个时间序列的周期性、不可预测性及波动性
28	临近点的绝对值之和
29	临近点的绝对值均值
30	时间序列的样本斜度

特征 31：能量比，即第 i 块子序列的平方和与整个序列的平方和之比。

特征 32：对时间序列分块后聚合，聚合后的值做线性回归，特征为由线性回归方程得到的标准差

特征 33：自回归模型的最大迟滞，自回归模型如下：

$$X_t = \varphi_0 + \sum_{i=1}^{k} \varphi_0 X_{t-i} + \varepsilon_i \tag{3.13}$$

特征 34：求最大迟滞的自相关性，公式如下：

$$\frac{1}{(n-1)\sigma^2}\sum_{t=1}^{n-l}(X_t-\mu)(X_{t+1}-\mu) \tag{3.14}$$

其中，X_i为时间序列，n 为时间序列的长度，σ^2及μ为方差和均值。

特征 35：平均二阶倒数：

$$\frac{1}{n}\sum_{i=1,\dots,n}\frac{1}{2}(x_{i+2}-2\cdot x_{i+1}+x_i) \tag{3.15}$$

特征 36：Ricker 小波的连续小波变换，其定义如下：

$$\frac{2}{\sqrt{3a}\pi^{\frac{1}{4}}}(1-\frac{x^2}{a^2})\exp(-\frac{x^2}{2a^2}) \tag{3.16}$$

其中 a 时小波函数的宽度参数。

特征 37：时间序列的复杂度，根据波峰、波谷计算。波峰波谷越多，序列越复杂，定义[52]如下：

$$\sqrt{\sum_{i=0}^{n-2lag}(x_{i-}x_{i+1})^2} \tag{3.17}$$

特征 38：衡量时间序列数据的非线性，定义[53]如下：

$$\frac{1}{n-2lag}\sum_{i=0}^{n-2lag} x_{i+2\cdot lag}{}^{2}\cdot x_{i+lag}\cdot x_i \tag{3.18}$$

3.4　本章小结

本章的主要工作及创新点如下：

(1) 基于 PCA 算法，设计实现了一种基于 AETA 电磁信号的前兆观测方法，并根据临震异常特点提取临震特征；

(2) 基于 Baer 算子，设计实现了一种基于 AETA 地声信号的前兆观测方法，并可作为临震特征输入下游模型；

(3) 选取了 38 种基于时间序列的特征提取方法及其对应的 76 种 AETA 特征。

第四章　基于分类模型的临震预测模型研究

地震成因是地震学科中的一个重要课题，目前有大陆漂移学说、板块构造学说等，但国际上还没有一致的定论[54]。研究中心认为，地震发生前后存在一个能量聚集及释放的过程，即会发生能量场、物理场的变化，这种变化可以作为前兆用于地震预测。但具体的某一前兆对地震预测是否具有指导作用，则需要度量前兆与地震之间的关系，再做进一步判断。在地震发震机理明晰之前，探究观测数据与地震之间的联系，有助于的地震预测这一课题的发展，而随着捕捉到的地震数据不断累积，地震预测模型也不断会趋于完善。

4.1　样本集的建立

地震预测是指根据对地震规律的认识，预测未来地震的时间、地震和震级，按照时间长短，可划分为长期预测、中期预测、短期预测和临震预测[4]。本文所研究的临震预测模型，用于对某地数日内，在较小范围内可能发生的破坏性地震做出预测。

本节分为两个部分：第一部分明确样本集中正负样本的划分标准；第二部分讨论样本不均衡问题及其解决办法。

4.1.1　地震预测三要素的选取

时间、地点和震级称为地震预测的“三要素”[4]。

地震震级是衡量地震大小的一种度量，根据地震时释放的能量来划分，释放的能量越多，震级越高。根据震级大小的分类情况，地震可分为弱震（震级＜3 级），有感地震（3 级≤震级≤4.5 级），中强震（4.5 级＜震级＜6 级）及强震（震级≥6 级），如表 4.1 所示。

表 4.1　破坏性地震划分标准

破坏程度	震级
弱震	＜3
有感地震	[3, 4.5]
中强震	(4.5, 6)
强震	≥6

破坏性地震一般是指震级大于 5 级，造成一定的人员伤亡和建筑物破坏或造成重

大的人员伤亡和建筑物破坏的地震灾害[57-58]。破坏性地震多为中强震及强震，会对人民的生产活动造成一定的影响，因而本文将震级预测抽象成两类:一类为发生破坏性地震，即 5 级及 5 级以上地震；一类为未发生破坏性地震，即未发生地震或发生地震但震级小于 5。

1971 年，傅承义先生在《地震战线》上发表文章《关于地震发生的几点认识》，提出“红肿理论”，认为“在一个较大地震（例如 Ms>4.5）发生之前，地壳上层在很大的地区内都已经起了变化，并不局限于岩层断裂的地区；断层不过是最后的爆发点而已。地震过程就仿佛人身上长疮一样，在一大片红肿的地方，疮口的面积只占一个很小的比例。在地震过程中，地面上的‘红肿’区是很大的，远远超过余震所限制的震源区。在这个‘红肿’区上，随处都可能发出地震的前兆。”[59]从这个观点出发，地震异常信号不应只局限在地震发生的时刻，可以作为前兆异常的地区也不应只限制在震源区。

由第三章中对震前异常分析可以看到，PCAETA 及 AETA-Baer 异常往往在地震前 3-5 日开始出现，因此，本文将模型的时间期限设定为 5 日，预测未来 5 天内的地震情况。

在分析实践中发现，出现明显电磁及地声异常的台站，往往距震源 200km 范围内。且对于临震预测而言，预报范围过大，会导致居民不必要的恐慌，并不具备实际意义。因此，本文将模型的地点限定为距台站 200km 范围内。

4.1.2　样本不均衡问题及其解决办法

上述分类中负样本无疑是充足的，而正样本相比是不足的，存在着样本不均衡问题，也叫作不平衡的类分布（imbalanced class distribution）[60]。

通常来说，数据集中不同类的样本数量往往具有一定的差异，小的差异是常见且无关紧要的，但在一些场景中，如欺诈性交易的数据集（多数的“非欺诈”类，少数的“欺诈”类）、客户流失数据集（多数的“未取消订阅”，少数的“取消订阅”），由于场景特点，类的样本数量之间会出现严重的不平衡，使得分类学习算法结果偏向大多数类，而在少数类实例中存在较高的错误分类率。

针对这样的问题，学者们提出了许多解决方案，可分为两大类，一类是数据抽样[61-63]，即对训练实例进行修改，以产生一个平衡的类分布，使分类器以类似于标准分类的方式执行；另一类是修改算法[64-66]，使算法模型更加适应样本不平衡问题。

其中前者即数据抽样，可分为三种做法[67]：

(1) 过采样：通过复制某些实例或从现有实例创建新实例来创建原始数据集的副本。过采样的代表性算法 SMOTE，它的工作原理是从少数类中创建合成样本，而不是创建副本。该算法选择两个或多个相似的实例（使用距离度量），并在该数据集内随机选择实例内的属性，构建新样本。

表 4.2　EasyEnsemble 算法

EasyEnsemble 算法

输入：少数类样本 P

多数类样本 N

多数类样本的子集数 T

训练 addaboostH_i的迭代次数s_i

输出：集合

$$\mathrm{H(x)} = \mathrm{sgn}(\sum_{i=1}^{T}\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x) - \sum_{i=1}^{T}\theta_i)$$

procedure: EasyEnsemble(P, N, T, s_i)

for i = 1 to T

从 N 中随机选择构成样本子集 Ni，使得$|N_i| = |\mathrm{P}|$

从 P+Ni 训练H_i，H_i由s_i个弱分类器$h_{i,j}$组成，$h_{i,j}$权重为$a_{i,j}$

集合H_i具有阈值θ_i

endfor

end procedure

(2) 欠采样：通过消除实例（通常是大多数类实例）创建原始数据集的子集。对于欠采样而言，如果简单地随机丢失样本，会损失一部分信息，因此为了解决这个问题，可使用欠采样的代表性算法 EasyEnsemble，EasyEnsemble 将多数类实例分组为大小与少数类相同的相等子集，令每个子集与少数类分别通过分类器进行处理，最后将结果集成在一起。EasyEnsemble 是一种集成式方法，它的每个子集都做了欠采样，但总体信息没有丢失，其算法流程如表 4.2 所示。

(3) 结合上述两种方案的 hybrids 方法。

分类中可能出现的一个问题是样本量小[68]。这一问题与“缺乏信息”有关，导致归纳算法没有足够的数据来对样本的分布进行概括。在高维数据（即大量特征）的存在下，这个问题会更加严峻。

样本不均衡问题对小样本集影响更大。在这种情况下，少数类可能表现得很差，因为学习该数据的模型的样本过少，会导致过拟合。因此，相比于大样本条件下的数据不平衡，小样本条件下的数据不平衡更具复杂性[69]。当然，训练数据中少数类样本质量的好坏也很重要。AETA 设备在全国布设 200 余套，但发生的较大震级的地震数量仍然较少，记录地震异常的数据占比也会比较小，当描述异常数据的特征维数较高时，这个

问题更为严峻。

4.2　特征选择

特征选择是预测模型中的一个重要环节[70]，它有助于减少原始特征中不相关或冗余信息的干扰，压缩输入数据的维数，保持预测模型的准确性。特征选择的另一个优势是，它在传统的数据驱动方法和物理模型分析之间架起了一座桥梁。因为这些特征都具有明确的物理意义，所以对筛选出的强特的分析，可以为今后地震探测物理研究提供一些启示。

4.2.1　特征选择方法

在机器学习的早期阶段，特征选择高度依赖于领域知识。在信息论和机器学习技术的基础上，开发了各种特征选择方法，并将其分为两类：筛选方法和包装方法[71]。

过滤方法的思想是根据一定的规则对所有特征进行排序，人工设定阈值来选择具有代表性的特征。这些方法一般在数据预处理阶段采用，与训练阶段无关，从而保证了较高的识别率和较低的计算复杂度。

包裹方法围绕不同的特征组合进行包装，并通过验证集上的学习算法的性能评估它们的有效性，验证集与过滤特征方法不同。通过在学习和更新阶段的评估，针对特定类型的学习算法，对包裹方法选择的特征进行了专门的优化。

本文叙述的特征选择方法如 Relief、Gini 系数、Kullback-Leibler 散度等都是筛选方法的典型示例，根据静态标准和阈值选择可能有用的“功能”。

Relief 算法通过特征区分相邻样本的能力来衡量特征的显著性。如果同一类样本之间，某特征的距离较大，则该特征不太有用，权重较低；相反，如果不同类的样本之间，该特征的距离较大，则说明该特征有用，权重较高。特征权重代表该特征的分类能力，权重越大，能力越强。

Gini 系数在经济学中，是统计离散度的一种度量，用来表示一个国家居民的收入或财富分布，但在数据挖掘领域，也可用于特征选择。在这里，本文使用 GI 来度量特征的分类能力。GI 表示样本中所有样本对，特征距离的均值：

$$GI = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \tag{4.1}$$

其中x_i及x_j代表样本对的特征值，n代表样本数量，$\bar{x}$代表所有样本的平均特征值。当所有样本的特征值相同，GI 值最小，为 0；理论上 GI 的最大值为 1，此时该表达式是一个极限不等式。当 GI 值较小时，代表该特征为不同的样本提供了相似的信息，以致样本的区分度不高。因此，可以将特征根据 GI 值进行排序，GI 值较高的特征可能含有更

有用的信息。

Kullback-Leibler 散度也称为相对熵，它用于衡量一个概率分布相对于另一个概率分布（期望概率分布）的偏离程度。假设 P(x)和 Q(x)代表两个离散概率分布，则 P(x)和 Q(x)的 KL 散度为：

$$D_{KL}(P||Q) = -\sum_{i=1}^{n} P(x_i) log \frac{Q(x_i)}{P(x_i)} \tag{4.2}$$

其中x_i为随机变量 x 的第 i 个元素，该度量表示该特征由 P(x)分布转化为 Q(x)分布时的熵减，因此 KL 散度越大的特征，越应该被丢弃，以避免冗余信息。

SLR（Sparse Logistic Regression，稀疏逻辑回归）是一种利用 L1 范数正则化的包裹式特征选择算法。L1 范数可以看作 L0 范数的最佳逼近，采用 L1 范数作为正则化项，可以使参数的绝对值之和变小。为保证性能良好，训练集数量需随弱参数量呈对数增长。SLR 将 L1 范数正则化添加到损失函数中，使得不太有用的属性具有较小的权重。因此，在逻辑回归过程中，重要特征可以具备较高权重。

4.2.2 特征选择实验及结果

实验台站选择九寨沟防震减灾局在 2017 年 7 月 4 日至 2017 年 12 月 1 日的数据，距离该台站 200km，5 日内若发生 5 级或 5 级以上的地震，则该日标记为正样本；否则标记为负样本。

本文使用第三节中描述的 78 种特征（包含两种人工特征，及 38 种基于时间序列的特征提取方法，在电磁及地声信号中提取的特征），应用三种特征选择方法（Relief、Gini 指数和 SLR）来评估每个特征对模型的潜在贡献。每种特征选择方法都会为这些特征打一个分数，以评估它们的潜在重要性。本文将低方差及每个特征选择指标下性能最差的 25%个特征丢弃，最终保留 25 个特征。详细信息汇总在表 4.3 中。

从表 4.3 可以看出，经验特征比时间序列的统计特征具有更高的选择率(100%)，这表明基于人类经验的特征在一维时间序列检测任务中可能具有一些优势。在四个不同领域（时域、频域、能量场、其它）的特征中，时域产生的特征贡献最大，它们在所有特征中所占的比例很大，在特征选择方法中也具有更高的选择倾向。

4.3 分类器的选择

分类在现代的许多场景中是一个常见的任务，本文将临震预测模型简化为一个二分类问题，从训练集中学习到的模型或分布函数即为分类器。由于本场景为小样本问题，因此，使用决策树或支持向量机这种参数较少的模型作为分类器会具有更好的表现。

表 4.3　特征描述

物理意义	信号类型	领域
PCAETA	电磁	人工特征
AETA-Baer	地声	人工特征
自回归系数	电磁	频域
最小重复记录检验	地声	时域
基于修正的 Fisher-Pearson 矩统计量的峰度	电磁	其它
时序数据非线性度量	电磁	其它
绝对能量值	电磁	能量场
扩展迪基-福勒检验（ADF 检验）	电磁	频域
基于分块时序聚合值的线性回归	电磁	其它
线性回归分析	电磁	其它
Ricker 小波分析	电磁	频域
时序数据区间内描述统计量	地声	其它
最大值位置	电磁	时域
重复记录检验	电磁	时域
一阶差分绝对和	地声	能量场
一阶差分绝对和	电磁	能量场
最小值位置	电磁	时域
最大值位置	地声	时域
均值下的最长连续子序列长度	电磁	时域
最小值位置	地声	时域
低于均值个数	电磁	时域
高于均值个数	电磁	时域
绝对傅里叶变换的谱统计量	地声	频域
最大值最近位置	地声	时域
最大值记录重复检验	地声	时域

4.3.1　决策树

决策树是一种经典的分类方法[72]，顾名思义，它是一种树形结构，其自顶向下的

生长过程其实就是模型对样本实例进行分类的过程。决策树无论是在分类问题或是回归问题中，都是特别有吸引力的模型类型，主要有以下三个原因：

(1) 具备良好的可解释性：它们具有一个直观的结构，生成的模型很容易被人类理解；

(2) 参数少：决策树是非参数模型，不需要人工干预，因此非常适合探索性知识发现；

(3) 可伸缩性好：当样本数量膨胀时，性能不会急剧下降。

a. 决策树的生成

决策树由结点及有向边组成，结点又可分为内部节点和叶子结点：内部结点表示的是满足某个特征或属性具体规则时的实例集合，叶子结点表示落在此结点的实例集合全部属于某个类，无需再分。

从根结点开始，决策树根据某种规则对样本的特征集合进行计算，选取分类效果最好的特征对当前实例集合进行划分，这样就将实例分配到了下一级结点。当结点为叶结点或者满足设定条件，则决策树停止生长。最后根据实例所在叶结点中多数实例对应的类别进行类别划分。图 4.1 是一个决策树的示意图。

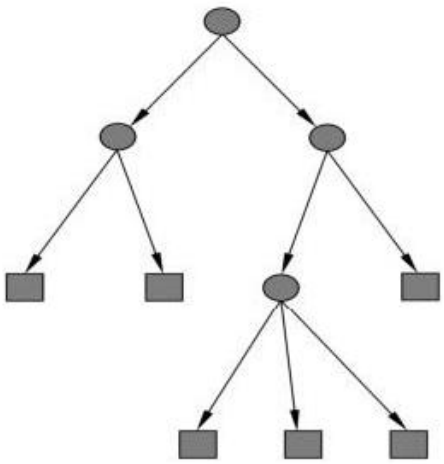


图 4.1　决策树模型

实际上，决策树中每一条从根结点到叶结点的有向路径都可以看成是一条对实例样本进行类别划分的规则；该有向路径上的每个内部结点对应着某个具体特征及其取值范围；叶结点的类别是由该结点中多数实例的类别决定的。用 if-then 规则来理解决策树，可以令实例的分类过程清晰可追溯，有助于理解特征，使分类过程具有可解释性。

特征选择是决策树生成过程中必要的一步，它的功能是选取能够对当前训练数据进行最佳划分的特征。在这个步骤中，规则的选择是非常重要的，有时甚至很难实现自

动化，所选特征的顺序将直接影响决策树的大小和性能。

为了精确地定义信息增益，这里需要定义一个信息理论中常用的度量，称为熵，它表征了任意示例集合的纯度。

给定一个集合 S，输出集 r，及输出所对应的的概率p_i (i = 1,2, … , n)，那么集合 S 相对于该分布的熵为：

$$H = -\sum_{i=1}^{r} p_i \times \log p_i \tag{4.3}$$

在所有涉及熵的计算中，定义$0 \log 0 = 0$。注意，如果集合 S 中所有样本都属于同一类，那么熵为 0；当集合 S 中每个类的样本数量相等时，熵为 1（最大值）；当集合 S 中类的样本数目不相等时，熵介于 0 和 1 之间。

在特征生成的过程中，如果用n_i来表示类c_i的样本个数，用n来表示数据集中所有的样本数，那么该数据集的经验熵为：

$$H_c \triangleq -\sum_{i=1}^{m} \frac{n_i}{n} \times \log_2 \frac{n_i}{n} \tag{4.4}$$

经验熵代表样本集中该类的纯度，用n_i^j来表示类c_j中对应特征为 j 的样本的个数，那么该特征相对于样本集的经验条件熵为：

$$H_{C|T} \triangleq -\sum_{j=1}^{p} \frac{n^j}{n} \sum_{i=1}^{m} \frac{n_i^j}{n^j} \times \log_2 \frac{n_i^j}{n} \tag{4.5}$$

信息增益为：

$$I_C^T \triangleq H_c - H_{C|T} \tag{4.6}$$

测试结果熵为：

$$H_T \triangleq -\sum_{i=1}^{p} \frac{n^j}{n} \times \log_2 \frac{n^j}{n} \tag{4.7}$$

Gini 值为：

$$G = 1 - \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{m} \frac{(n_i^j)^2}{n^j} \tag{4.8}$$

特征选择具有多种准则可供选择。CART 分类回归树使用 Gini 系数作为度量；ID3 算法使用信息增益，即特征 X 的先验信息对数据集 Y 的分类过程具有的指示作用大小以及对不确定度降低的贡献量，作为度量。但信息增益具有其固有缺点：当特征的取值越多，该特征越容易被选，使用 ULg 可以避免这个问题，它的定义如下：

$$gain(T) \triangleq \frac{2 \times I_C^T}{H_C + H_T} \tag{4.9}$$

b. 决策树的剪枝

剪枝对于决策树是非常重要的。有两种方法来解决这个问题：要么前瞻性地决定

何时停止一棵树的生长，即预剪枝；要么回顾性地对一棵完整决策树进行修剪，即后剪枝。通常，修剪是一个自上而下的过程，根据某些限定条件，每个节点可能替换为子树或者叶节点，下面给出一个通用模板的伪代码，如表 4.4 所示：

预剪枝方法建立了一些规则，阻止似乎不能提高树的分类能力的分支停止生长，这些规则的示例如下：

(1) 所有的样本属于同一类；

(2) 结点中的样本虽然不属于同一类，但具有相同的特征向量；

(3) 结点中样本数量小于某个阈值；

(4) 扩张对系统性能的增益太低。

表 4.4　决策树的剪枝

决策树的剪枝
输入：决策树 T
函数 f(T, m) 返回 T 的泛化误差，m 为样本数目
foreach 结点 j 自上而下（自下而上）遍历：
寻找 T’，使得 f(T, m)最小，T’需满足以下任一条件：
将节点 j 替换为叶 1 后的当前树。
将节点 j 替换为叶 0 后的当前树
将节点 j 替换为其左子树后的当前树
将节点 j 替换为其右子树后的当前树
当前树
T:= T’

4.3.2　支持向量机

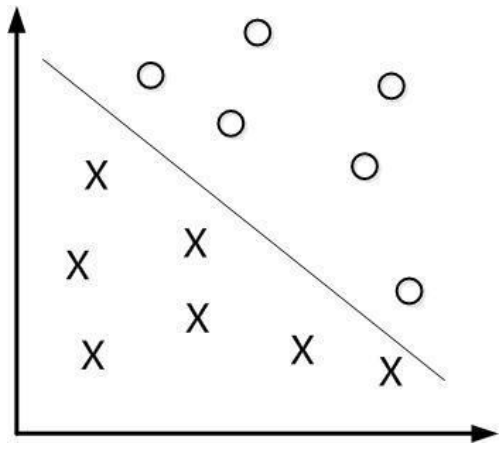


图 4.2　支持向量机

SVM[73]（support vector machines，支持向量机）可能是最流行和最受关注的机器学习算法之一，它是一种二分类模型，最初用来解决线性可分问题，但使用核技巧，使得支持向量机可以将数据从低维空间映射到高维空间取解决，由此成为一个非线性分类器。

a. 线性可分支持向量机

SVM 的基本思想非常直观，即寻找一个超平面，可以将多维空间内的正负样本正确分开，并且其几何间隔最大，如图 4.2 所示。几何间隔是指数据点到该超平面的距离。

设超平面为：$f(x) = \omega \cdot x + b$，其中ω及 b 为超平面的参数，则样本点$P(x_i, y_i)$到该平面的几何距离为：$d = \frac{|\omega \cdot x_i + b|}{||\omega||}$。若该超平面可以将正负样本完全分开，则对于任一样本，都有：

$$\text{distance}(x_i, y_i) = y_i \cdot \frac{|\omega \cdot x_i + b|}{||\omega||} > 0 \tag{4.10}$$

找到对所有样本点几何间隔最大的超平面，意味着离超平面最近的样本点（支持向量）距离该平面的距离也是最大的，用数学表示为：

$$\min_{\omega,b} \left[\min_{x_i} \frac{y_i(\omega \cdot x_i + b)}{||\omega||} \right] \tag{4.11}$$

等比缩放ω, b，使得$y_i(\omega \cdot x_i + b)$最小的点 $P(x_i, y_i)$有$y_i(\omega \cdot x_i + b) = 1$，则对于其他样本点有$y_i(\omega \cdot x_i + b) > 1$。这里，$y_i(\omega \cdot x_i + b)$称为函数间隔。

对样本点进行限制后，式 XX 内层表达式为$\frac{1}{||\omega||}$，该问题即转化为

$$\min_{\omega,b} \frac{1}{||\omega||}, \text{ s.t. } y_i(\omega \cdot x_i + b) \geq 1 \tag{4.12}$$

由于在数学问题中，凸优化问题更好求解，故可继续将问题转化为：

$$\min_{\omega,b} \frac{1}{2}||\omega||^2, \text{ s.t. } y_i(\omega \cdot x_i + b) \geq 1 \tag{4.13}$$

上式可用拉格朗日对偶法求解，定义拉格朗日函数：

$$L(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\omega \cdot x_i + b)) \tag{4.14}$$

对ω, b求导，令其值为 0，求解可得：

$$\begin{cases} \omega = \sum_{i=1}^{m} \alpha_i x_i y_i \\ 0 = \sum_{i=1}^{m} \alpha_i y_i \end{cases} \tag{4.15}$$

带回式(4.13)，可得到对偶问题：

$$max\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j {x_i}^T x_j$$

$$\text{s.t. } \textstyle\sum_{i=1}^{m}\alpha_i y_i, \tag{4.16}$$

$$\alpha_i \geq 0, \quad \mathrm{i}=1,2,3,\dots,m$$

求解对偶问题，可解出α，进而解出ω和b。

$$\begin{aligned}\mathrm{f(x)} &= \omega\cdot\varphi(x)+\mathrm{b}\\ &= \textstyle\sum_{i=1}^{m}\alpha_i y_i {x_i}^T x+b\end{aligned} \tag{4.17}$$

上述过程需满足 KKT 条件：

$$\begin{cases}\alpha_i \geq 0\\ y_i\mathrm{f(x)}-1\geq 0\\ \alpha_i(y_i\mathrm{f(x)}-1)=0\end{cases} \tag{4.18}$$

该对偶问题的二次规划问题可使用 SMO 算法进行求解，同时，对偶问题引入核函数的概念。

b. 核函数

核函数将线性不可分的低维数据投射到高维空间中求解，使得 SVM 转化为一个非线性分类器，使用$\varphi(x)$来表示x映射到高维空间后的特征向量，则超平面表示为：

$$\mathrm{f(x)}=\omega\cdot\varphi(x)+\mathrm{b} \tag{4.19}$$

目标函数表示为：

$$\min_{\omega,b}\frac{1}{2}\left|\left|\omega\right|\right|^2, \text{s.t. } y_i(\omega\cdot\varphi(x)+b)\geq 1 \tag{4.20}$$

对偶问题为：

$$max\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\varphi(x_i)^T\varphi(x_j)$$

$$\text{s.t.}\textstyle\sum_{i=1}^{m}\alpha_i y_i, \tag{4.21}$$

$$\alpha_i\geq 0, \quad \mathrm{i}=1,2,3,\dots,m$$

由于直接计算$\varphi(x_i)^T\varphi(x_j)$比较困难，设想一个函数$\kappa(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$，使得$x_i$和$x_j$在高维空间的内积，可通过$\kappa(\cdot,\cdot)$来计算，此时对偶问题可简化为：

$$max\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(x_i,x_j)$$

$$\text{s.t.}\sum_{i=1}^{m}\alpha_i y_i, \tag{4.22}$$

$$\alpha_i \geq 0,\ \mathrm{i}=1,2,3,\ldots,m$$

求解后得到

$$\begin{aligned}\mathrm{f(x)} &= \alpha_i y_i \mathrm{x}+\mathrm{b}\\ &= \alpha_i y_i \varphi(x_i)^T\varphi(x_j)+\mathrm{b}\\ &= \alpha_i y_i \kappa\left(x_i,x_j\right)+\mathrm{b}\end{aligned} \tag{4.23}$$

这里κ(·,·)即为 SVM 的核函数，其所构造的非线性支持向量机结构如图 4.3 所示。表 4.5 给出几种常用的核函数[47]。其中线性核和高斯核在实际使用中比较常用：当特征数量足够多的，使得特征不需要映射到更高维的特征空间时，可以使用线性核，它的参数少，速度快；当使用线性核无法得到满意的效果时，可以使用高斯核，它的参数多，结果很耗时，但由于线性核是高斯核的简并情况，通过大量的参数调试后，高斯核的结果往往比线性核要好。

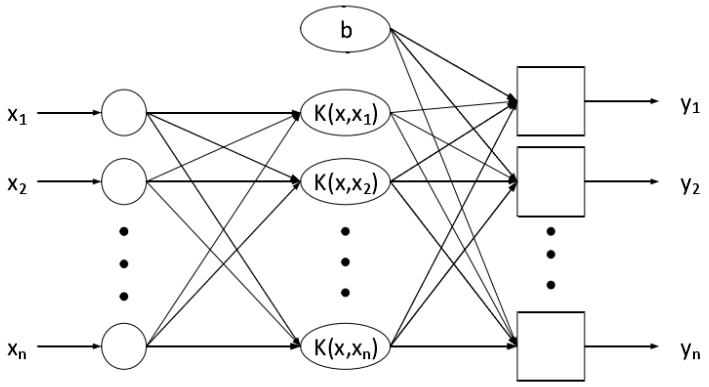


图 4.3　支持向量机的非线性结构

SVM 具有以下特点：

(1) SVM 不适用于大数据集，因为 SVM 使用二次规划的方法来求解，该方法涉及到 m 阶矩阵的运算，如果 m 很大，将消耗大量配置；

(2) SVM 的决策函数由距离超平面最近的支持向量来决定，而与其它非支持向量无关。因此，非支持向量的增减不会对 SVM 的分类能力造成影响，这个特性使得 SVM 具备很好的泛化能力；

(3) SVM 自带 L2 正则，因此在小样本集上往往可以取得比其它分类器更好的结果。

表 4.5　常用的核函数

名称	表达式	参数
线性核	$\kappa(x_i, x_j) = {x_i}^T x_j$	
多项式核	$\kappa(x_i, x_j) = ({x_i}^T x_j)^d$	$d \geq 0$为多项式的次数
高斯核	$\kappa(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$	$\sigma > 0$为高斯核的带宽(width)
拉普拉斯核	$\kappa(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|}{\sigma})$	$\sigma > 0$
Sigmoid 核	$\kappa(x_i, x_j) = \tanh(\beta {x_i}^T x_j + \theta)$	tanh 为双曲面函数，$\beta > 0$，$\theta < 0$

4.3.3　其它对比算法

除了上述两种经典的机器学习模型，为了探索分类器选取的有效性，本文选取了几种其它经典的机器学习模型作为对比，其中包括 RF（Random Forest，随机森林）、GBDT（Gradient Boosting Decison Tree，梯度提升树）和 XGBoost（eXtreme Gradient Boosting，极端梯度提升）。

1. RF

随机森林是一种集成算法（Ensemble Learning），它属于 Bagging 类型，通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能。其可以取得不错成绩，主要归功于“随机”和“森林”，一个使它具有抗过拟合能力，一个使它更加精准。

2. GBDT

梯度提升树是一种迭代的决策树算法，它基于集成学习中的 boosting 思想，每次迭代都在减少残差的梯度方向新建立一颗决策树，迭代多少次就会生成多少颗决策树。它与随机森林一样，都是通过多棵决策树的有机组合获得最终结果。不同的是，随机森林属于并行关系，而梯度提升树更像是串行关系。

3. XGBoost

和 GBDT 一样，XGBoost 也是一种提升树模型。不同的是，GBDT 只支持 CART 作为基分类器，而 XGBoost 还支持线性分类器，并且在使用线性分类的时候可以使用 L1，L2 正则化；另外 XGBoost 使用二阶泰勒展开，特征子采样以及通过与排序，使得模型可以进行并行化处理，使得模型的优化速度大大增加。

4.4　模型评估方法及参数优化

4.4.1　性能评估方法

性能评估是模型开发工作中不可或缺的一部分，它有助于找到最佳模型以及模型在未来的运作情况[74]。在数据科学中，使用训练数据来进行模型效果评估是不可接受的，因此这样很容易会引起过拟合或欠拟合。数据科学中有一些常用的评价方法，这里介绍 3 种：Holdout 检验，K-Fold 交叉验证和自助法[75]。为了避免过拟合，这些方法都是用测试集来评估模型性能。

1. Holdout 检验

在此方法中，将大型数据集随机分为三个部分：

(1) 训练集：用于构建预测模型的子集；

(2) 验证集：用于评估训练集构建的模型性能的子集。它提供了一个测试平台，用于微调模型参数，并选择性能最佳的模型。并非所有的建模算法都需要验证集；

(3) 测试集：未被选取过得数据所构成的子集，用于评估模型的泛化能力。如果训练集对模型的效果评估好于测试集，那么可能是因为发生了过拟合。

Holdout 检验适用于大数据集，因为评估结果与分组方式（如训练集和测试集的正负样本比）有很大的关系，使得测试评估效果与训练评估效果之间可能具有额外的偏差。

2. K-Fold 交叉验证

在一些场景中，数据是稀缺的，如果不想在验证时“浪费”数据，可以使用 K-Fold 交叉验证，这项技术旨在在不浪费太多数据的情况下准确估计真实误差。

在 K-Fold 交叉验证中，原始训练集被划分为 m/K 大小的 K 个子集（为简单起见，假设 m/K 是整数）。对于每一个子集，该算法使用其他子集的并集进行训练，然后用该子集估计其输出的误差。最后，所有这些误差的平均值就是对真实误差的估计。特殊情况 K=m，其中 m 是样本数，称为“留一法”。

K-Fold 交叉验证通常用于模型选择（或参数调整），一旦选择了最佳参数，算法将在整个训练集上使用该参数进行重新训练。下面给出了模型选择的 K-Fold 交叉验证的伪代码，参照表 4.6。输入为训练集 S、参数值候选集θ、子训练集个数 K，及算法 A；输出为最优参数θ^*，及在此参数下算法 A 的输出。

3. 自助法

自助法是一种重采样技术，通过从原始数据集中有放回地抽取样本（可能会有重复），来生成与原始数据集大小相同的新数据集，作为训练集；而流程中没有被选中的样本，则作为测试集，用作验证。

相比于 Holdout 检验及 K-Fold 交叉验证方法，自助法不需要单独划分一部分数据

用来测试，因而更加充分了利用原始样本的信息，因而在小样本数据集上具有更好的表现。然而，使用重复的样本进行训练出的模型，与原始数据的分布有所不同，因而也引入了偏差，因此，如果数据量足够，应优先使用 Holdout 检验及 K-Fold 交叉验证方法。

综上所述，本文采用 K-Fold 交叉验证法制作训练集和测试集，其中 K 为地震发生的次数，每次训练时，选取一个地震附近的数据作为测试集，其它数据为训练集。

表 4.6　K-Fold 交叉验证

K-Fold 交叉验证

输入：

训练集 S=$(x_1, y_{1),\dots,}(y_m, y_m)$

参数值候选集Θ

分类器 A

整数 k

将 S 划分为S_1，S_2，…，S_k

foreach $\theta \epsilon \Theta$

for $i = 1 \dots k$

$h_{i,\theta} = A(S \backslash S_i; \theta)$

$\text{error}(\theta) = \frac{1}{k}\sum_{i=1}^{k} L_{S_i}(h_{i,\theta})$

输出：

$\theta^* = \text{argmin}_\theta[\text{error}(\theta)]$

$h_{\theta^*} = A(S; \theta^*)$

4.4.2　性能评估指标

模型选择及超参数、核函数等参数调优，是机器学习研究和应用的重要步骤。通常，调参的目标是找到其中的最优解，为此需要构建评估指标，以对不同的机器学习方法、参数、模型之间进行选择。不同的场景往往需要不同的指标来进行评估，因为大多数的评估指标都有其固有局限性，只能反映模型在某一方面的性能，对分类模型算法的改进也是在一定方面的改进。这里给出几种常用的评估指标及其局限性。

1. 混淆矩阵

混淆矩阵是一个表，如表 4.7 所示，通常用于描述一组测试数据的分类模型的性能，其中所有值是已知的。除了 AUC 意外所有的测量值都快可以通过这个表中四个参

数来计算，下面来谈谈这四个参数：

TP 及 TN 是预测正确的结果；FN 和 FP 是希望最大程度地减少误报和漏报。这些术语可能有些晦涩，下面是每个术语的意义：

- 真阳性（TP）：正确预测的正样本，这意味着实际结果是正样本，预测结果也是正样本。
- 真阴性（TN）：正确预测的负样本，这意味着实际结果是负样本，预测结果也是负样本。
- 误报（FP）：当实际结果为负样本时，预测结果为正样本。
- 漏报（FN）：当实际结果为正样本时，预测结果为负样本。

当测试结果和实际结果相符时，真阳性及真阴性非负；当测试结果与实际结果相悖时，误报和漏报非负。一旦理解了这四个参数，就可以据此计算出准确率，精度，召回率，和 F1 分数。

表 4.7　混淆矩阵

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

2. 准确率

准确率是最直观的性能指标，它是正确预测的样本数占总样本数的比率。

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.24}$$

一般来说，准确率越高，模型效果越好。准确率具有其固有缺点，即忽略了误报和漏报性能，因此只有当数据集对称时，准确率才是一个比较好的评估指标。

3. 精确率与召回率

精确率是正确预测的正样本占预测为正样本的比率；召回率是正确预测的正样本占实际正样本的比率，其表达式如下：

$$Precision = \frac{TP}{TP+FP} \tag{4.25}$$

$$Recall = \frac{TN}{TN+FN} \tag{4.26}$$

4. F1-measure

F1-measure 是精确率和召回率的加权平均值。该指标考虑了误报和漏报，直觉上它不想准确性那么容易理解，但 F1-measure 往往比准确性更有用，特别是类分布不均

匀的场景。如果误报和漏报具有相同的结果，那么准确性更好；如果误报和漏报的结果十分不同，那么最好同时看看准确率和召回率。F1-measure 的表达式如下：

$$F1 = \frac{2\times precision\times recall}{precision+recall} \tag{4.27}$$

5. P-R 曲线

P-R 曲线和 ROC 曲线一样，都是二分类模型的评价工具，允许在一个范围内可视化功能。P-R 曲线在机器学习领域中的应用越来越广泛，尤其在非平衡数据集中具有良好的表现。一个典型的 P-R 曲线图如图 4.4[76]所示，目标是让模型位于右上角。除了观察 P-R 曲线图外，可以通过计算 P-R 曲线下的面积（AUCPR）及计算$precision = recall$时的阈值(BEP)作为性能的度量。

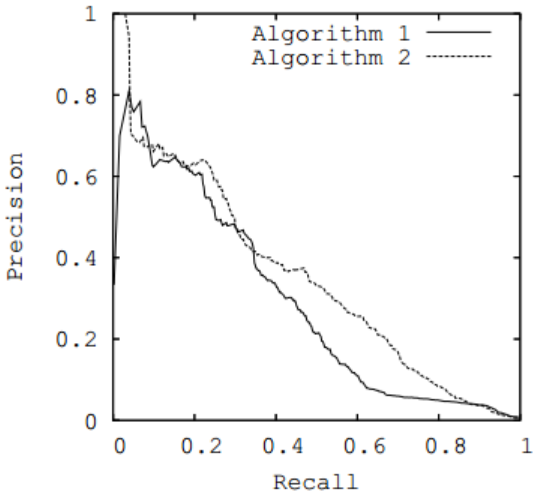


图 4.4　P-R 曲线示例图

6. ROC 曲线与 AUC 值

ROC 曲线起源于 20 世纪 40 年代的声呐，用于测量从噪声中监测到的声呐信号的程度，目前在机器学习领域有着广泛的应用。ROC 曲线和 P-R 曲线一样，也是二分类模型的评价工具，允许在一个范围内可视化功能。

ROC 曲线绘制了不同分类阈值下的 TPR 和 FPR。降低分类阈值将更多的样本被划分为正样本，从而增加假阳性和真阳性。图 4.5[76]显示了典型的 ROC 曲线。FPR 和 TPR 的计算方法分别为

$$FPR = \frac{FP}{N} \tag{4.28}$$

$$TPR = \frac{TP}{P} \tag{4.29}$$

为了计算 ROC 曲线上的点，可以用不同的分类阈值多次评估一个模型，但这样做

效率很低。幸运的是，有一种高效的、基于排序的算法可以提供这些信息，称为 AUC。

AUC 代表“ROC 曲线下的面积”，也就是说，AUC 测量从（0,0）到（1,1）的 ROC 曲线下的整个二维面积。它提供了对所有可能的分类阈值的性能的汇总度量，其取值范围为 0 到 1，预测 100%错误的模型的 AUC 为 0.0；预测 100%正确的模型的 AUC 为 1.0。

AUC 具有两个明显的特性：

(1) AUC 衡量排名的好坏，而不是绝对值；一种解释 AUC 的方法是，模型将随机正例比随机负例排列得更高的概率。

(2) AUC 不需要对阈值进行选取，直接衡量模型预测的质量。

这两个特性使得 AUC 在推荐、计算广告等领域极受欢迎，然而在一些场景，如垃圾邮件监测(希望将误报最小化)，AUC 不是一个有用的度量。

通常，当每个类的数量大致相等时，应使用 ROC 曲线，当类中存在较大的不平衡时，应使用 P-R 曲线；但当数据变化时，ROC 曲线具有更好的鲁棒性。

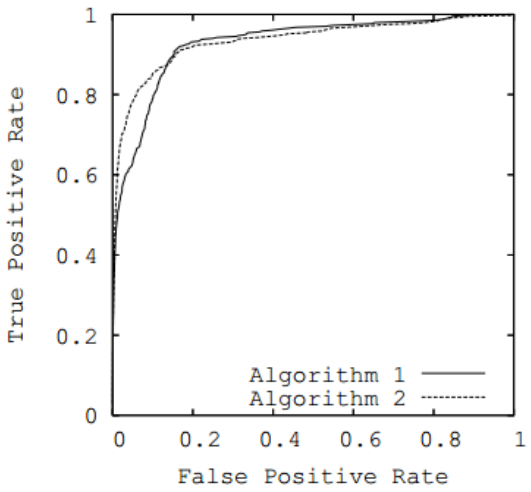


图 4.5　ROC 曲线示例图

4.4.3　超参数调优

在机器学习中，超参数优化或调整是为模型选择一组最优超参数的问题。超参数是一个用来控制学习过程的参数，它来自先验分布，对模型的性能产生重大影响[77]。常用的超参数调优技术包括：网格搜索，随机搜索，及贝叶斯优化。

1. 网格搜索

网格搜索是一种传统的超参数调优方法，网格搜索尝试为各种超参数的每个组合构建一个模型并评估，以选择提供最佳性能的超参数配置。网格搜索需要预先指定超参数的范围及补偿，以便配置网格。在实际应用中，使用者通常从参数值之间具有相对

较大步长的网格开始，然后在最佳配置下扩展或使网格更精细，并继续重现搜索网格。此过程成为手动网格搜索。

网格搜索是一种低效的方法。假设有 n 组超参数，每组超参数有两个值，那么配置总数为2^n，陷入维度诅咒。因此，仅在少量配置上进行网格搜索是可行的。幸运的是，网格搜索中超参数值的选择不依赖于先前训练的结果，因而很容易并行化，且不会影响性能，这使得在足够的计算能力下网格搜索变得可行。

2. 随机搜索

随机搜索使用超参数的随机组合来找到构建模型的最佳方案。在随机搜索中，作者展示了一个令人惊讶的结果：通过随机选择超参数网络，可以获得与网格搜索类似的性能。

随机搜索非常简单和有效，因此许多从业者认为它是调整超参数更优先的方法。像网格搜索一样，超参数值的选择不依赖于先前训练的结果，它可以很容易地实现并行，但和网格搜索相比，它的实验次数要少得多，而性能相当。

3. 贝叶斯优化

区别于网格搜索及随机搜索方法，贝叶斯优化使用上一次迭代的结果来改进下一次训练，属于 SMBO 算法。贝叶斯优化使用高斯过程（GP）函数来获得后验函数以基于先验函数进行预测。随着观测次数的增加，替代函数得到改善，算法对参数空间中哪些区域值得探索，哪些区域不值得探索变得更加确定。

相比于网格搜索和随机搜索，贝叶斯优化在大的样本集中具有更好的表现，但它只能处理数值型参数。

4.5 模型的反演实验

4.5.1 样本集构建

AETA 终端软件成熟于 2016 年 12 月。AETA 台站在 2017 年 1-6 月在社会各界支持下，先后在台湾、四川、云南、河北完成了布设。在九寨沟地震后的 8 月和 9 月份又进行了大批量的布设。截止 2018 年 11 月 22 日，全国共安装 196 个台站。其中 2017 年 7 月 1 日前安装的台站有 89 个，2016 年 12 月 31 日前的台站只有 10 个，2017 年 10 月 1 日前的台站有 115 个。

根据中国地震台网数据统计，2018 年位于 AETA 台站布设密集的川滇地区及其附近共发生 5 次破坏性地震，且皆发生在下半年，本文以这 5 次破坏性地震事件，及其 200km 内 2018 年 7 月 1 日-2018 年 12 月 31 日的 AETA 数据为基础，构建样本集。样本集中正样本容量为 220，正负样本比约为 1:18，存在一定程度正负样本不平衡的情况，因此对负样本进行降采样，使得正负样本比为 1:1。

4.5.2　反演效果

本文选用 4.3 节中各分类器对样本集进行训练，并且将结果与其它模型进行对比，结果如表 4.8 所示。

由表中可以看出，决策树及 SVM 的查准率及 AUC 指标表现都明显优于 RF、GBDT、XGBoost 模型。该结果是符合预期的，因为本场景的样本容量较小，相比于参数较少的决策树、SVM 模型，RF、GBDT、XGBoost 模型需要“学习”的参数较多，因而需要更多的数据对模型进行拟合，因此，它们在小样本的场景中容易发生过拟合现象；而决策树和 SVM 模型需要“学习”的参数较少，因此它们在该场景中表现会更好。

表 4.8　各地震预测模型对比的表现

模型 / 指标	基于 AETA 数据的临震预测模型					前兆研究	危险理论	
	RF	**GBDT**	**XGBoost**	**决策树**	**SVM**		**DCA**	**BP**
查准率	0.540	0.506	0.515	**0.660**	0.608	0.437	0.52	0.5
查全率	0.772	**0.910**	0.773	0.750	0.705	—	0.58	0.56
AUC	0.556	0.511	0.523	**0.682**	0.625	—	—	—

另外，作者也将本模型与其它人的工作进行对比。任等[14]的工作是通过虎皮鹦鹉跳跃现象、次声波、引潮力共振、地应力等前兆信号，对某一地区（如北京地区、日本北海道），未来 3-7 日，震级为 5.5 级以上的地震进行预测，该方法查准率达到 43.7%；甘等[78]同样使用 AETA 数据构建地震预测模型，对距台站 100km 范围内，未来 30 日，3.0 级以上的地震进行预测，其基于 DCA 算法模型的查准率及查全率为 0.52、0.58；基于 BP 神经网络模型的查准率及查全率为 0.5、0.56。

与根据前兆信号进行经验性预测的实验相比，本文构建的临震预测模型，预测的时间跨度更短，地域范围更小，震级相似且查准率更高；而与同样使用 AETA 数据的模型相比，本文构建的临震预测模型预测时间跨度更短，且查准率及查全率具有更好的表现。

4.6　本章小结

本章的主要工作及创新点如下：

(1) 针对数据特点，提出样本不均衡问题的对应解决方案，构建样本集；

(2) 对经验特征及常规特征进行特征选择，筛去潜在贡献小的特征，保证下游模型泛化能力更强，减少过拟合。

(3) 构建地震预测模型，并进行反演实验，查准率可达 0.66，查全率可达 0.75。

第五章　总结和展望

5.1　总结

地震预测是一项意义重大且极具挑战的工作，无数研究者为此付出了大量的心血与努力。本文主要以基于多分量地震监测系统 AETA 进行临震特征提取及预测模型搭建为目标，在阅读大量相关文献及调研最新技术后，提出并设计了 78 种临震特征及临震预测模型，为地震预测研究工作提供了一些有益的研究思路及探索结果。下面总结一下本文前四章的工作：

(5) 基于 AETA 数据的预处理

在 AETA 系统的实际运行过程中，不可避免地会出现断电、断网的问题，导致数据会出现缺失、突跳情况。本文针对 AETA 电磁扰动数据和地声数据的特点，对于缺失数据，分别使用临近数据及线性插值的方法进行填补；对于突跳数据，截取突跳部分，再视为缺失数据进行补全。数据预处理工作为后续的模型搭建奠定了基础。

(6) 基于 AETA 数据的特征提取

本文从人工经验特征和通用统计特征两个方面来进行特征提取。既考虑了在实践过程中，对临震信号进行观察获取的经验总结，又充分利用了前人总结的通用特征所带来的便利性，尽可能为下游模型提供丰富的候选项。

在人工经验特征方面，本文分析了电磁扰动的临震异常现象，提出基于主成分分析方法的 PCAETA 特征；分析了地声的临震异常现象，提出基于 Baer 算子的 AETA-Bear 特征。

在通用统计特征方面，本文基于时域、能量、频率特点，选用 76 种统计特征，从多个维度对 AETA 数据进行描述，为下游的特征筛选工作提供丰富的候选项。

(7) 特征选择

本文提出的原始特征达到 78 种，这些特征中可能包含了一些冗余特征，特征选择可以减少原始特征中不相关或冗余信息的干扰，压缩输入数据的维数，有助于提高预测模型的性能，保持预测模型的准确性。因此，本文使用 Gini 系数、SLR 等特征选择方法对原始特征进行打分，最终选定贡献度最大的 25 个特征，进入下游的分类器进行临震预测。

(8) 临震预测模型

本文提出并实现了一种临震预测模型，对距台站 200km 范围 5 日内的破坏性地震（震级≥Ms5.0）进行预测。由于场景具有小样本的特点，分类器选用了决策树及支持

向量机。在回溯实验中，本文选取 2018 年 7 月 1 日至 2018 年 12 月 31 日，地理位置位于川滇地区及其附近的 5 次破坏性地震及距震源 200km 内的 AETA 数据为基础，构建样本集。实验结果表明，构建的临震预测模型查准率可达 0.66，查全率可达 0.75，对地震预测问题的解决具有一定的意义。

5.2　展望

本文从 AETA 监测系统的电磁及地声数据出发，对一定时间及范围内的破坏性地震进行预测。所构建的模型具备一定的效果，但是和精准的临震预测这一目标，还相差甚远，结合这几年的心得体会，未来可在这些方面深入研究：

(1) 基于单个台站构建样本集，存在样本小的问题；基于所有台站构建样本集，存在地理结构特征淹没的问题。因此，以小范围的数个台站为单位，如龙门山地带，进行研究，可能会有更多的灵感和进展。

(2) 本文所构建的 78 个特征（包括 2 个人工特征及 76 个统计特征）具有明确的物理意义，包含了信号在时域、频域、能量场等方面的信息，对位于断裂带及非断裂带的台站进行特征分析，将有助于断裂带的探测工作。

参考文献

[1] Geller R J. Earthquake prediction: a critical review[J]. Geophysical Journal of the Royal Astronomical Society, 2010, 131(3):425-450.

[2] 张勇, 陈运泰, 许力生. 2009 年 4 月 6 日意大利拉奎拉（L′Aquila） MW6.3 地震的破裂过程——视震源时间函数方法与直接波形反演方法比较[J]. 地球物理学报, 2010, 53(6):1428-1439.

[3] 李岩峰, 王广余. 2001—2010 年全球有人员死亡的灾害性地震综述[J]. 国际地震动态, 2011, 2011(11):16-20.

[4] 赵永红, 杨家英, 惠红军等. 地震预测方法Ⅰ:综述[J]. 地球物理学进展, 2014, 29(1):129-140.

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

[6] Jordan T H , Chen Y T , Gasparini P , et al. Operational Earthquake Forecasting: State of Knowledge and Guidelines for Utilization[J]. Translated World Seismology, 2011, 54(4):315-391.

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

[8] Geller R J, Jackson D D, Yan Y K, et al. Earthquakes cannot be predicted[J]. Science, 1997, 275(5306):1616-1616.

[9] Asim K M, Idris A, Iqbal T, et al. Seismic indicators based earthquake predictor system using Genetic Programming and AdaBoost classification[J]. SOIL DYNAMICS AND EARTHQUAKE ENGINEERING -SOUTHAMPTON-, 2018.

[10] Richter C F. Seismicity of the Earth[J]. Nature, 1970, 225(5228):170-170.

[11] 郭星, 潘华, 李金臣等. 一种基于经验分布的大地震复发概率计算方法[J]. 地震学报, 2018, v.40(04):110-122.

[12] 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.

[13] Raleigh C B, Sieh K, Sykes L R, et al. Forecasting southern california earthquakes[J]. Science, 1982, 217(4565):1097-1104.

[14] 任振球, 李均之, 曾小苹. 大地震临震预测的研究进展[J]. 地学前缘, 2001, 8(2).

[15] 钱尚玮, 周永明. 临震异常判别的统计分析[J]. 地球物理学报, 1980(2):105-117.

[16] 张敏灵, 周志华. 机器学习专题编者按[J]. 中国科学: 信息科学, 2017, 47: 1443

[17] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245):255-260.

[18] Panakkat A, Adeli H. NEURAL NETWORK MODELS FOR EARTHQUAKE MAGNITUDE PREDICTION USING MULTIPLE SEISMICITY INDICATORS[J]. International Journal of Neural Systems, 2007, 17(01):13-33.

[19] 聂红林, 袁孝, 胡伍生等. 基于 BP 神经网络技术的区域短期地震预测模型研究[J]. 现代测绘, 2012, 35(2):3-5.

[20] 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.

[21] Asim K M, F. Martínez-Álvarez, Basit A, et al. Earthquake magnitude prediction in Hindukush region using machine learning techniques[J]. Natural Hazards, 2017, 85(1):471-486.

[22] 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-.

[23] 石耀霖, 孙云强, 罗纲等. 关于我国地震数值预报路线图的设想——汶川地震十周年反思[J]. 科学通报, 2018, 63(19):1865-1881.

[24] Zhang G M , Liu J , Shi Y L . An scientific evaluation of annual earthquake prediction ability[J]. Acta Seismologica Sinica, 2002, 15(5):550-558.

[25] 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.

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

[27] Molchanov O A, Mazhaeva O A, Protopopov M L. Electromagnetic VLF radiation of seismic origin observed on the interkosmos-24 satellite[J]. Geomagnetizm I Aeronomiya, 1992, 32(6): 128-137

[28] Matsushima M, Honkura Y, Oshiman N, et al. Seismoelectromagnetic effect associated with the Izmit earthquake and its aftershocks[J]. Bulletin of the Seismological Society of America, 2002, 92(1): 350-360

[29] Karakelian D, Klemperer S L, Fraser-Smith A C, et al. Ultra-low frequency electromagnetic measurements associated with the 1998 MW 5.1 San Juan Bautista, California earthquake and implications for mechanisms of electromagnetic earthquake precursors[J]. Tectonophysics, 2002, 359(1-2): 65-79

[30] 汤吉, 詹艳, 王立凤等. 5 月 12 日汶川 8．0 级地震强余震观测的电磁同震效应[J]. 地震地质, 2008(3):739-745.

[31] 汤吉, 詹艳, 王立凤等. 汶川地震强余震的电磁同震效应[J]. 地球物理学报, 2010, 53(3): 526-534.

[32] 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): 0511-0520.

[33] 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.

[34] Bleier T, Freund F. Earthquake [earthquake warning systems] [J]. IEEE Spectrum, 1996, 42(12):22-27.

[35] 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.

[36] 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.

[37] Bermbach D, Wittern E, Tai S. Data Preprocessing[M]// Cloud Service Benchmarking. 2017.

[38] Roy A, Cruz R M O, Sabourin R, et al. A Study on combining Dynamic Selection and Data Preprocessing for Imbalance Learning[J]. Neurocomputing, 2018, S0925231218300936.

[39] Baer M , Kradolfer U . An Automatic phase picker for local and teleseismic events[J]. Bulletin of the Seismological Society of America, 1987, 77(4):1437-1445.

[40] Wan X, Wang W, Liu J, et al. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range[J]. BMC Medical Research Methodology, 2014, 14(1).

[41] Fuying Z, Yun W, Jian L, et al. STUDY ON METHOD OF DETECTING IONOSPHERIC TEC ANOMALY BEFORE EARTHQUAKE[J]. Journal of Geodesy and Geodynamics, 2009, 29(3):50-54.

[42] Marchetti D, Akhoondzadeh M. Analysis of Swarm satellites data showing seismo-ionospheric anomalies around the time of the strong Mexico (M w =8.2) earthquake of 08 September 2017[J]. Advances in Space Research, 2018, S0273117718303776.

[43] Liu J, Jiang C, Deng C, et al. Vertical ionosonde net and its data application in southwestern China[J]. Acta Seismologica Sinica, 2016.

[44] Zhang Z, Wang Y, Chen Y, et al. Crustal structure across Longmenshan fault belt from passive source seismic profiling[J]. Geophysical Research Letters, 2009, 36(17):1397-1413.

[45] Xu J, Yan W, Tang W. Risk Factors of Post-traumatic Stress and Depressive Disorders in Longmenshan Adolescents After the 2013 Lushan Earthquake[J]. Community Mental Health Journal, 2018(12):1-10.

[46] Matsushita R, Imanishi K, Ohtani M, et al. Seismic Potential Around the Northeastern Edge of the Longmenshan Fault Zone as Inferred from Seismological Observations[J]. Pure and Applied Geophysics, 2019(2).

[47] Yan Z K, Wang X B, Yong L I, et al. Coupling between the dynamic processes at depth and geologic processes on the surface of the Longmenshan thrust belt[J]. Chinese Journal of Geophysics, 2017.

[48] 王新安,雍珊珊,黄继攀等. 基于 AETA 监测数据的地震预测研究[J].北京大学学报(自然科学版), 2019, 55(02):209-214.

[49] Chang X, Zou B, Guo J, et al. One sliding PCA method to detect ionospheric anomalies before strong Earthquakes: Cases study of Qinghai, Honshu, Hotan and Nepal earthquakes[J]. Advances in Space Research, 2017, 59(8):2058-2070.

[50] 邹斌, 郭金运, 常晓涛等. 基于主成分分析与滑动四分位法的震前 TEC 异常探测对比分析[J]. 全球定位系统, 2016, 41(4):63-69.

[51] 张小红, 任晓东, 吴风波等. 震前电离层 TEC 异常探测新方法[J]. 地球物理学报, 2013, 56(2):441-449.

[52] Batista G E A P A, Keogh E J, Tataw O M, et al. CID: An efficient complexity-invariant distance for time series[J]. Data Mining and Knowledge Discovery, 2013, 28(3).

[53] Schreiber T. Discrimination power of measures for nonlinearity in a time series[J]. Phys. Rev. E, 1997, 55.

[54] 郭增建, 秦保燕. 地震成因和地震预报[M]. 地震出版社, 1991.

[55] 叶洪, 马瑾, 汪一鹏等. 从破裂模拟实验探讨破坏性地震发震条件的一些初步成果[J]. 地质科学, 1973, 8(1):48-55.

[56] 张益民. 破坏性地震人员被困及伤害特点对地震应急训练的启示[J]. 中国西部科技, 2011, 10(23):5-6.

[57] Fernández A, García S, Herrera F. Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution[J]. Hybrid Artificial Intelligent Systems, 2011, 6678:1-10.

[58] Feng H U, Wang L, Zhou Y. An Oversampling Method for Imbalance Data Based on Three-Way Decision Model[J]. Acta Electronica Sinica, 2018, 46(1):135-144.

[59] 傅承义, 陈运泰, 陈顒. 我国的震源物理研究[M]. 1979.

[60] Yang X, Kuang Q, Zhang W, et al. AMDO: an Over-Sampling Technique for Multi-Class Imbalanced Problems[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, PP(99):1672-1685.

[61] Wang Y, Gan W, Wu W, et al. Dynamic Curriculum Learning for Imbalanced Data Classification[J]. 2019.

[62] Nikpour B, Nezamabadi-Pour H. HTSS: a hyper-heuristic training set selection method for imbalanced data sets[J]. Iran Journal of Computer Science, 2018(3):1-20.

[63] Huang C, Li Y, Loy C C, et al. Deep Imbalanced Learning for Face Recognition and Attribute Prediction[J]. 2018.

[64] Jin X, Ling X, He C, et al. Dynamic classifier ensemble model for customer classification with imbalanced class distribution[J]. Expert Systems with Applications An International Journal, 2012, 39(3):3668-3675.

[65] Fernández A, García S, Herrera F. Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution[J]. Hybrid Artificial Intelligent Systems, 2011, 6678:1-10.

[66] Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance[J]. IEEE Trans.pattern Anal.mach.intell, 1997, 19(2):153-158.

[67] Wasikowski M, Chen X W. Combating the Small Sample Class Imbalance Problem Using Feature Selection[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(10):1388-1400.

[68] Lord D, Mirandamoreno L F. Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter of Poisson-Gamma Models for Modeling Motor Vehicle Crashes: A Bayesian Perspective[J]. Safety Science, 2008, 46(5):751-770.

[69] Pan S J, Qiang Y. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(10):1345-1359.

[70] Oaksford M, Chater N. Optimal data selection: revision, review, and reevaluation.[J]. Psychonomic Bulletin & Review, 2003, 10(2):289-318.

[71] Li J, Liu H. Challenges of Feature Selection for Big Data Analytics[M]. 2017.

[72] Quinlan J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1):81-106.

[73] Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters, 1999, 9(3):293-300.

[74] Mason R L, Gunst R F, Hess J L. Model Assessment[J]. 2003.

[75] Vehtari A, Ojanen J. A survey of Bayesian predictive methods for model assessment, selection and comparison[J]. Statistics Surveys, 2012, 6(1):1-1.

[76] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves[C]//Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 233-240.

[77] Toal D J J, Bressloff N W, Keane A J. Kriging Hyperparameter Tuning Strategies[J]. Aiaa Journal, 2012, 46(5):1240-1252.

[78] 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.