版权声明

摘要

地震作为一种具有强大破坏力的自然灾害，对人类社会造成巨大的伤害。时至今日，国内外的地震预测研究与地震预报实践仍然处于较低水平，特别是最具实用价值的短临预测，还无法对三要素进行有效预测，无法满足社会的需求。为了对地震预测这一世界难题进行求解，北京大学深圳研究生院集成微系统重点实验室研制了多分量地震监测系统 AETA，给地震预测带来了数据基础，同时也对如何处理 AETA 数据并基于 AETA 数据进行地震预测提出了新的要求。本文所做的主要工作如下：

提出了 AETA 分层数据处理的方法，将 AETA 传感器监测的时序数据转化为台站当天的地震风险标签。在各个分层中分别完成如下工作：（1）在数据预处理层完成对断电数据和缺失值的处理，并实现了 AETA 数据采样率的统一；（2）在特征提取层构造了 13 个统计特征和 1 个分形维数特征，丰富了 AETA 数据的信息表达；（3）在异常检测层分别使用滑动四分位距和DBSCAN密度聚类完成了对一维数据和多维数据的异常提取，在保留数据时序特性的同时对不同台站的同一特征进行了“归一化”；（4）基于地震三要素生成样本标签，把地震风险模型的学习目标确立为未来数天台站附近有无发震风险的二分类问题，基于特征提取层、异常检测层的数据，使用随机森林算法建立了地震风险模型。运用 AETA 分层数据处理方法处理了 36592 个验证样本，AUC 为0.696;而在2018年10月31日西昌Ms5.1地震相关的450个样本中,AUC可达0.822。

提出了对地震三要素按震中、发震时间、震级顺序进行独立预测的方法。（1）在震中预测方面，构建了风险及位置特征集把发震风险和地理位置关联起来，并通过均值漂移聚类进行震中预测。运用所得的震中模型分析 2018 年 6 月 29 日平武 Ms4.0 地震，在震前两天和发震当天所计算出的预测震中到实际震中的距离分别是 80.4KM、72.4KM、61.3KM。（2）在已知地点、锁定时间的条件下，对未来地震震级进行预测。基于地震风险特征和震例，训练了三个对固定地点、固定时间（未来三天、五天、十天）可能出现的地震震级进行多分类预测的模型。在验证集上，忽略对无震类别的预测效果后，对未来三天出现 Ms3 至 Ms4.5 地震的查准率、查全率可达 0.745、0.667，而对未来五天、未来十天的地震震级预测查准率和查全率都低于 0.5。在 2018 年 10 月 31 日西昌 Ms5.1 地震这一具体震例上，尽管三个模型在该震例的样本中平均准确率只有 0.43，但综合三个模型的预测结果，在震前两天可以锁定该地震的震级为 Ms4.5 至 Ms6。

关键词：地震预测，多分量地震监测系统 AETA，随机森林

AETA Hierarchical Data Processing and Research on Earthquake Prediction Model

Yuanhao Feng (Microelectronics and Solid-State Electronics)
Directed by Xin'an Wang

ABSTRACT

As a natural disaster with strong destructive power, the earthquake has caused tremendous damage to human society. Today, earthquake prediction research and earthquake prediction practice at home and abroad are still at a low level, especially the short-term prediction with the most practical value. It is still unable to effectively predict the three factors and cannot meet the needs of society. In order to solve the world problem of earthquake prediction, the Key Laboratory of Integrated Microsystems of Peking University Shenzhen Graduate School developed the multi-component seismic monitoring system AETA, which brought the data foundation to the earthquake prediction work, and also how to process the AETA data based on The data for earthquake prediction puts forward new requirements.

This paper proposes a method of AETA hierarchical data processing, which converts the time series data monitored by the AETA sensor into the seismic risk label of the station. The following work is done in each layer: (1) Processing the power-off data and missing values in the data pre-processing layer, and realizing the unification of the AETA data sampling rate; (2) 13 structures are constructed in the feature extraction layer. Statistical features and a fractal dimension feature enrich the information expression of single data of AETA data; (3) Complete the anomaly of one-dimensional data and multi-dimensional data by using the sliding interquartile range and DBSCAN density clustering in the anomaly extraction layer respectively. Extracting, while preserving the data timing characteristics, the same feature of different stations is "normalized"; (4) Based on the seismic three elements to generate sample tags, the learning objectives of the seismic risk model are defined as near the station in the next few days. Based on the feature extraction and anomaly extraction data, the seismic risk model was established using the random forest algorithm. Using AETA hierarchical data processing method, 36,592 verification samples were processed, and the

average AUC was 0.696. In the 450 samples related to the Xichang Ms5.1 earthquake in Sichuan on October 31, 2018, the AUC reached 0.822.

A method for independent prediction of the three elements of the earthquake according to the epicenter, time and magnitude order is proposed. (1) In the epicenter prediction, the risk and location feature set is constructed to correlate the earthquake risk with the geographical location, and the epicenter prediction is performed by mean drift clustering. Using the obtained epicenter model to study the Sichuan Pingwu Ms4.0 earthquake on June 29, 2018, the distances from the predicted epicenter to the actual epicenter calculated on the two days before the earthquake and on the day of the earthquake were 80.4KM, 72.4KM, and 61.3KM, respectively. (2) Under the condition of locking time, predict the magnitude of future earthquakes. Based on the seismic risk characteristics and earthquake cases, three models for multi-classification prediction of earthquake magnitudes that may occur in fixed locations and fixed time in the next three days, five days, and ten days are trained. On the verification set, after neglecting the prediction effect on the non-seismic category, the precision and the recall rate of the Ms3 to Ms4.5 earthquakes in the next three days can reach 0.745 and 0.667, and for the next five days and the next ten days. The earthquake magnitude prediction precision and recall rate are both below 0.5. On the specific earthquake case of the Xichang Ms5.1 earthquake in Xichang, Sichuan Province on October 31, 2018, although the average accuracy of the three models in the sample of the earthquake case was only 0.43, the prediction results of the three models were combined before the earthquake. The day can roughly lock the magnitude of the earthquake to Ms4.5 to Ms6.

KEY WORDS: Earthquake prediction, AETA, Random forest

目录

第一章　绪论

1.1　课题背景及研究意义

地震是一种具有强大破坏力的自然灾害，地震灾害造成的死亡人数占各类自然灾害造成的死亡人数总数的 54%，堪称群灾之首[1] 。我国东临环太平洋地震带、西达欧亚地震带，大部分国土都受到太平洋板块、印度板块和菲律宾海板块的挤压，川滇地区、西北地区、华北地区、东南沿海地区以及台湾省都是地震多发区域。总体而言，我国地震具有活动频度高、破坏性地震多、地震范围广等显著特征，是全球大陆地震灾害最严重的国家之一。

地震带来的巨大伤害也激起了人类对地震进行研究的热情，自 19 世纪 70 年代现代地震学创立以来，预测地震、减轻地震灾害这些问题一直被热切地讨论着。现代地震学的主要研究方向有：测震、震源物理和前兆信号研究等[2]。直到 20 世纪 60 年代，信息科学日新月异的发展，催生了一大批测震和地震监测仪器，随着数据的积累，现代地震学的前景变得更加明朗。

除了人们在震前所感知的“宏观”前兆，如地震云、动物异常等现象外，前兆信号研究主要依靠地震监测仪器来探测地震发生前后各类地震信号的异常变化。地震监测的前兆信号研究方法可分为地球物理方法、大地形变测量和地球化学等方法[3]。常见的前兆信号有：裂纹或岩石的摩擦特性变化、重力场变化、磁场变化、电场变化、地下电阻率变化、地下流体流动、地应变加速或地壳形变、地下水化学成分变化、大气化学成分变化。通过研究这些信号与地震的关系，从而进一步探索预测地震的可能性。

为了探索并解决地震监测预测这一世界难题，北京大学深圳研究生院集成微系统重点实验室研制了多分量地震监测系统 AETA（下称 AETA 系统）。AETA 系统能够实时采集地表的电磁扰动信号和地声信号，并且通过网络传输到云服务器。同时，AETA 系统满足大区域密集布设中高灵敏度、低成本和易布设的需求，现已在云南、四川、西藏、河北、北京、广东等地区密集布设[4]。

自 2017 年 6 月稳定以来，AETA 系统已经积累了近 2 年的丰富数据。结合地震科学与数据科学，对 AETA 系统的监测数据进行科学的分析、处理，并在此基础上探索和求解地震预测的三要素：发震时间、地点、震级，成为当下的重要课题。此课题的研究，能充分挖掘 AETA 系统监测数据的丰富内涵，并尝试从数据科学的角度对地震三要素的预测进行求解，为地震预测这一重大的科学问题进行有益的探索。

1.2 国内外研究状况

1.2.1 地震预测方法研究现状

地震预测的主要目标，是预测未来地震的时间、地点和震级。由于地震发生的机理复杂且时间上随机，准确预测地震三要素是公认的世界难题。根据预测时间跨度分为长期预测、中期预测、短期预测和临震预测。相较于中长期预测，短临预测更具实际意义。而国内外地震预测研究与地震预报实践总体水平仍然不高，特别是短临预测，还远远无法满足社会的需求[3] 。AETA 系统所监测的电磁扰动和地声信号是当今短临预测中重要的前兆信号，有望应用于短临预测。

目前，短临预测可分为基于历史地震信息和基于地震前兆信号两大类。基于历史地震信息的短临预测，大多是通过以往地震的时空分布以及相应震级来推测未来地震的三要素。而基于地震前兆信号的短临预测，则是在数据层面上从各类的地震前兆信号中提取出有用的特征值，并根据所得特征值推测未来地震的三要素。

1. 基于历史地震信息

基于历史地震信息进行短临预测的方法包括传统的概率统计和近年兴起的机器学习方法，两类方法本质上是寻找地震事件在时间、空间、能量这个三维特征空间上的分布[5,6]。目前，影响力最大、使用层面最广的是概率统计方法得出的古登堡-里克特规则(G-R 规则)和特征地震分布[7]。而机器学习方法则利用自身强大的非线性映射能力从更丰富的地震事件特征中寻找地震事件的时空能量分布，常见的地震事件特征如 G-R 规则的 a 值和 b 值、历史最大地震与当前地震震级差值和间隔时间、地震能量的平方根变化率等。

2. 基于地震前兆信号

随着信息技术的进步，各种精密的地震监测仪器的发明[8–11]，越来越的地震前兆信号也被发现，基于地震前兆信号的短临预测方法成为了研究的热点。常见的地震前兆信号，包括地应变加速或地面隆升、电磁体的变化、地下气液体化学成分变化等等[12–14]。基于不同的地震前兆信号，所采用的数据处理方法、所提取的特征值也各不相同。应力及形变类的观测手段，常关注于应力的大小变化、形变的类型，并通过数据模拟建立应变场、位移场来研究地震的三要素[15–17]，1987 年尹祥础等提出了加卸载响应比理论定量反映震源区介质的变形、损伤过程，并成功预测 1989 年美国加州 Ms7.1 地震[18]。对前兆信号的观察主要集中于原始波形的幅度、脉冲信号的频率以及电磁脉冲能量的大小，数据处理方法则以平均值法、中位值法、四分位距法、滑动时窗法等线性方法为主[19–21]。近年来，也有一些研究使用机器学习和时间序列方法对地震信号进行分析[22,23]，但这些方法大多关注单一的前兆信号，三要素的预测也基本依赖于信号提

取出的个别特征。而事实上，地震前兆信号的变化具有不确定性，现在还无法找到任何一种“确定性的地震前兆信号”，依靠单一信号、稀少的特征值对地震三要素进行预测，是很难具备符合要求的泛化能力的。

本文希望建立一个支持多种地震前兆信号的分层数据处理框架，并结合机器学习方法对地震预测模型进行研究，以提高地震预测的精度。

1.2.2　基于电磁扰动信号的地震研究现状

地震电磁学是地震前兆研究的一个重要分支。在孕震过程中，震源区及其附近常会发生不同程度的电磁扰动异常，且这些异常通常都会在震前或发震时出现[24–27]。与地质学、地球物理学和地球化学等多个其他研究手段相比，电磁扰动（电场、磁场变化）是一种较好捕捉的地震短临异常，电磁扰动异常在震前几天甚至几个小时常出现明显的异常变化，是短临地震研究中反映最灵敏的前兆信号之一。在许多震例中，地震电磁扰动的变化与地震震级呈正相关，震级越大，异常越明显[28]。根据地震前电磁扰动的变化幅度、异常持续时间等信息来研究地震的短临预报具有良好的研究前景。

地震电磁的观测方法主要有地基和天基两类。以国家地震电磁台网为代表的地基方式，在经过半个多世纪的攻坚，在全国布设了各类电磁观测台站，对特定区域的数据进行连续性的观测，积累了丰富的电磁数据，对数百个震例进行了分析，在震源区电磁扰动特征、地震电磁的形成机理、电磁信号与地震活动的相关性等领域有了长足的发展[29,30]。区别于地基方式的区域限制，以法国 DEMETER 卫星为代表的天基方式能从一个更宏观的区域观测电磁的变化，各国科学家基于卫星数据对地震期间电离层的扰动以及其耦合机制开展了相关研究，并在多次强震中观测到较为明显的电子浓度异常[31–35]，但扫描式的运行方式使它无法对特定区域进行连续性的观测，且在机理研究方面还有所欠缺。

在多次物理实验和实际震例中，发现地震孕育期间岩石的破裂会出现压电、压磁效应，这主要是在岩石产生破裂时，岩石晶格破坏后电位跳跃会辐射出电磁波信号[36–39]。刘君等在研究芦山 Ms 7.0 地震和岷县漳县 Ms 6.6 地震时，发现两次地震前 4 个台站原始波形出现了脉冲信号频繁突跳、极低频成分 PSD 增大等异常现象，同时电磁脉冲能量都出现了明显的增大[40]。范莹莹等发现 2008 年汶川 Ms 8.0 地震震中周围 8 个地电阻率台站在震前出现不同形态的异常变化，断裂带附近的地电场和电磁扰动出现波形畸变和能量增强；甚至沿龙门山断裂带南北 1300KM 外的河北电磁扰动台也出现自观测以来最大幅度的异常变化[41]。高曙德等发现 2008 年国内 5 次 Ms 6.0 以上地震都能在乌鲁木齐和通海 2 个地震台站找到对应的电磁异常，异常包括：电磁场自功率谱密度在震前出现明显的脉冲异常且异常幅度与震中距负相关、脉动异常具有时间上

的集丛现象等[42]。丁跃军等认为强震前电磁扰动异常的变换过程符合地震活动的发展规律，且电磁扰动异常的幅度、时间与震级和震中距有关[43]。

现有的电磁信号数据处理方法主要可分为四类：(1) 时域方法包括差值法、时空参考场法、空间互相关法等[44–47]；(2) 频域方法包括谐波分析法、频谱分析法、小波分析法[48–51]；(3) 特征转换类的方法有：转换函数法、感应矢量法、低点位移法[52–55]；(4) 统计学方法，如主成分分析法、分形分析法等[56–59]。

1.2.3 基于地声信号的地震研究现状

在诸多地震前兆中，地声作为一种直接获取地下信息的信号，具有很大的研究价值。地声信号包括大地震前的可听前兆地声以及中、小地震前地下传播的高频振动（广义前兆地声）。地震的孕育过程常伴随着地应力场的变化、地壳形变的慢速变化及蠕变，当局部地壳受压而发生形变或微破裂时，会出现地声，其频率范围从几赫兹到几千赫兹之间；而由地震纵波引起的岩石宏观破裂，也会在横波达到地面前产生可听前兆地声，可以理解为地震能量以声波的形式传递到空中；除此之外，在岩石破裂过程中，被高温或其他因素导致的地下电离气体，在接近地面处放电时也会产生地声[60]。综上，地声可以反映出地震孕育过程中的某些地球物理化学信息，是地震预报研究中的重要领域之一。

我国的地声观测研究最早可追溯到 1966 年邢台地震，一直到 1983 年后才陆续在山东莒县、云南洱源和四川江油等地建立了使用仪器记录可听地声的地声监测台网[61]。蒋锦昌等人从 1985 年开始探索地声与地震的关系，通过综合地声波在传播过程中的衰减特性及小白鼠体内 5-HT 代谢过程的研究，得出地声的信号优势频段[62]。1994 年，郑治真等从山东莒县和云南洱源的地声台网观测数据中证实了前兆性地声的存在，并认为微破裂是地声产生的主要机制[60]。2007 年我国基于 GPRS 网络建成了地声监测系统，从而开启了大范围的地声研究。近年来，陈维升等专注于地震前兆地声异常的研究，直观的从地声异常信号进行分析，对近几年世界上的大地震有着较好的短临预测[63,64]。

国外关于地声的记录最早出现在 20 世纪初，直到 1976 年 Hill 等在体波地动与地声关系的研究中，发现地声信号的成因是地震 P 波和 SV 波通过震动与空气耦合所产生的，这一发现大大促进了地声研究的进展[65]。2007 年，Lin 和 Langston 开展了声音与震源的研究，发现依据声音可经验性判定天然地震源的响应特性[66]。2010 年，Laštovička 等人通过在地震震中同时进行的次声、地震、磁场和电离层的观测，证明了地震运动的垂直分量在激发的次声波中的主导作用。2018 年，Shahar 等人通过对意大利中部地震的地声耦合信号研究，认为大气、地质环境会对地震地声的传播产生影响，

并推断弹性介质中的地震波在声学介质中的次声波传播的区域传播范围要比过往认为的更大[67]。

近年来，地声被越来越多的地震学家关注。我国现阶段地震监测台网已建立了一批数字地震观测网络试点，这批试点可全天候观测并实现了较好的自动化，但台网密度较小，观测点分布非常分散。同时，由于地声信号的频率范围广、信噪比低，对地声信号的研究处理还有很大的提升空间。

1.3　本文研究工作

为了探索并解决地震预测这一世界难题，北京大学深圳研究生院集成微系统重点实验室研制了多分量地震监测系统 AETA，并在地震频发的四川、云南、西藏、河北、北京、广东、台湾等地区密集布设。自 2017 年 6 月稳定运行以来，AETA 系统已经积累了近 17TB 的原始数据。

本文基于对 AETA 数据的理解和对地震三要素预测任务的拆分，提出了 AETA 分层数据处理框架，以充分挖掘 AETA 系统监测数据的丰富内涵。同时，从数据科学的角度构建了震中预测模型和震级预测模型。AETA 数据、AETA 分层数据处理框架、震中预测模型以及震级预测模型的具体内容及相互关系，如图 1.1 所示：

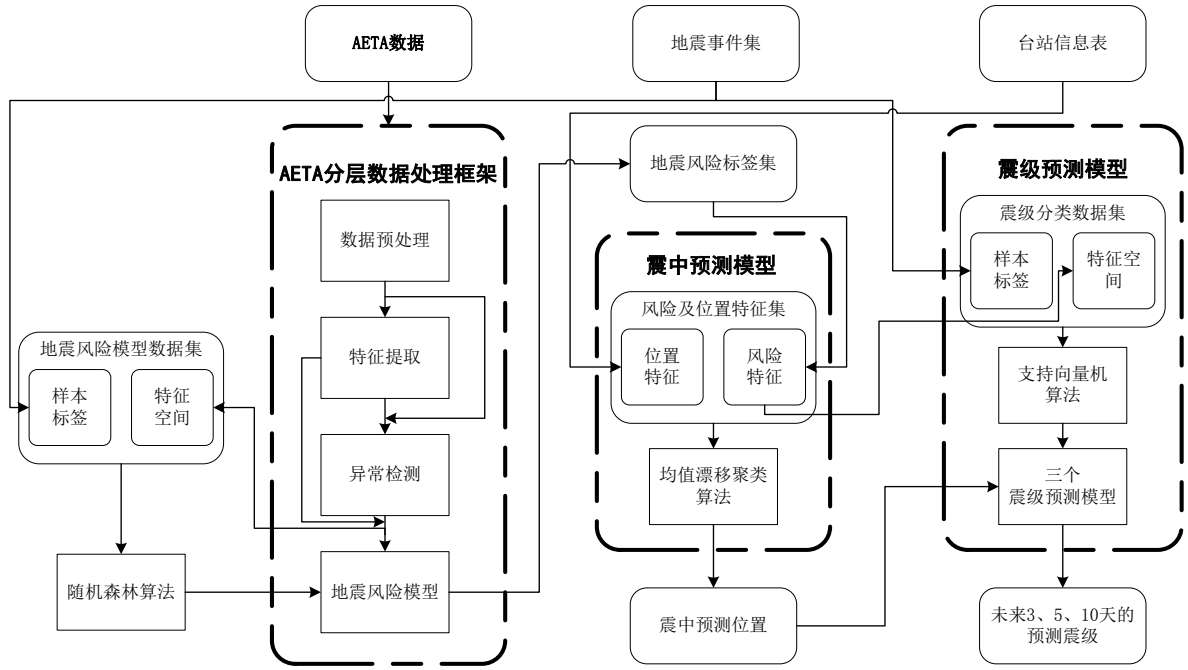


图 1.1　AETA 数据、AETA 分层数据处理框架、震中预测模型以及震级预测模型

在进行地震预测时，AETA 数据会先经过 AETA 分层数据处理框架，转化为各台

站当天的地震风险标签；此后，震中预测模型会在由地震风险标签及台站信息表构成的风险及位置特征集上，进行均值漂移聚类从而预测出震中的具体位置；最后，三个震级预测模型根据震中预测位置和周围台站的风险特征，预测未来 3 天、5 天、10 天可能出现的地震震级。

由图 1.1 可知，本文主要进行了以下三个主要工作：

1. 提出了 AETA 分层数据处理框架

AETA 分层数据处理框架包括数据预处理、特征提取、异常检测、地震风险模型四个层次。该框架的目标是将 AETA 传感器监测的时序数据转化为台站当天的地震风险标签，从而关联 AETA 数据和地震事件。数据预处理层的核心目标，是去除系统的“脏数据”并对时序数据进行规整， 从而提高 AETA 数据的数据质量并为之后的处理提供良好的数据基础；特征提取层的主要作用是从 AETA 时序数据中获取结构化数据，丰富 AETA 数据的信息表达，降低后续模型的收敛难度；异常检测层的主要作用是从特征数据中提取异常值或异常点，更加关注特征序列的趋势和异常波动的描述，解决特征的绝对数值不能很好地刻画地震事件的问题，并在保留时序数据的时序特点的同时在某种程度上对不同台站进行归一化；地震风险模型层的核心目标是训练一个可以描述特征提取层和异常检测层的结构化数据和地震三要素之间关系的模型。

2. 实现了 AETA 分层数据处理框架

在数据预处理层，完成对断电数据的自动化识别和清洗、对缺失值的分类填补、实现数据重采样模块以解决 AETA 数据传输采样率不一致的问题；在特征提取层，基于传统统计方法提取十二种统计特征、基于分形理论提取 AETA 电磁扰动数据的分形维数，从而获得十三种采样率为一天一个点的特征序列；在异常检测层，通过滑动四分位距法计算一维数据的异常值、通过 DBSCAN 密度聚类对对于多维数据进行异常点检测，从而获得多种采样率为一天一个点的异常序列；在地震风险模型层，基于地震三要素生成地震风险标签、结合特征序列及异常序列构建特征空间，在五个备选算法中挑选随机森林算法并完成对地震风险模型的构建。

3. 提出并实现了对地震三要素进行独立预测的方法

针对无法对地震三要素进行准确预测的现状，提出先预测震中，再锁定预测的未来天数，最后预测震级的预测方法。（1）在震中预测上，基于地震风险特征表现相似的台站在地理位置上应可聚为同簇的猜想，提出通过聚类进行震中预测的研究思路并实现了一个震中预测模型。本文首先基于各台站的地震风险标签和台站经纬度，构建风险及位置特征集。其次，根据经纬度的距离换算和地震风险特征的尺度换算，对这个地震风险-地理位置的高维特征空间进行特征变换。最后，通过均值漂移聚类的方法找出具有地震风险的簇中心，并把簇中心的地理信息提取出来，作为预测震中的具体

位置。（2）在时间和震级预测上，基于 AETA 分层数据处理框架生成的地震风险标签集能较好地描述台站附近的地震风险、震中预测模型输出的预测震中位置比较准确的两个前提下，提出一种可基于附近台站的地震风险推算该震中未来数天的可能震级的震级预测模型。本文首先基于地震事件和各台站的地震风险特征构建用于多分类震级预测的数据集；其次，为进一步明确三要素中的时间，通过支持向量机算法训练三个分别对三天内、五天内、十天内可能出现的地震震级进行多分类预测的震级预测模型。

1.4　论文组织架构

基于本文的研究内容，论文的组织结构如下：

第一章，绪论部分，首先分析了地震的危害以及地震预测的重要意义，并说明了对 AETA 数据进行处理并在此基础上进行地震预测研究的重要意义。其次，分析了国内外地震预测的研究现状，并详细说明了基于电磁扰动和地声的地震预测进展。最后介绍了本文的研究工作和论文组织架构。

第二章，分析 AETA 数据并实现 AETA 分层数据处理中的数据预处理层。首先系统介绍了 AETA 的项目背景、系统架构、观测探头、安装情况，并对 AETA 数据采集流程进行了全面梳理。其次，实现 AETA 分层数据处理中的数据预处理层，具体包括断电数据处理、缺失值处理和重新采样与频率转换。

第三章，提出并实现 AETA 分层数据处理框架中的特征提取层和异常检测层。在特征提取方面，研究并实现基于统计方法和分形理论对 AETA 数据进行特征提取的方法。基于统计方法提取了三大类统计特征：基本统计值、刻画地声震前异动、刻画 SRSS 波的日升日落特性；基于分形理论提取了 AETA 电磁数据的分形维数。在异常检测方面，采用滑动四分位距法计算一维数据的异常值、采用 DBSCAN 密度聚类对多维数据进行异常点提取。

第四章，提出并构建 AETA 分层数据处理框架中的地震风险模型。首先基于地震三要素生成地震风险标签，并基于特征提取层、异常检测层的输出结果构建了地震风险模型的特征空间。其次，通过分析算法的适用范围和地震风险模型的特征空间，从五个分类算法中选取了随机森林算法。最后，在模型训练过程中，使用 SMOTE 过采样解决了数据集样本不平衡的问题，结合经验和网格搜索法选取出合适的模型超参数，使用五折交叉验证训练了基于随机森林算法的地震风险模型，并对模型效果进行分析讨论。

第五章，研究并实现了震中预测模型和震级预测模型。针对震中预测模型，首先提出了台站间地震风险特征和地理位置在分布上存在相似性的猜想，并基于各台站的

地震风险标签和台站经纬度，构建了风险及位置特征集；其次，研究了均值漂移聚类算法并针对风险及位置特征集设计了具体的特征变换方法；最后，利用聚类算法从风险及位置特征集中对震中进行预测，并展示了在实际震例中的预测效果。针对震级预测模型，首先确立了模型的目标：对固定位置未来数天可能出现的地震震级进行预报。其次，基于地震事件和台站的地震风险特征，构建了多分类数据集。再次，利用主成分分析进行特征降维并通过支持向量机训练了三个震级预测模型，分别对未来三天内、未来五天内、未来十天内可能出现的地震震级进行多分类预测。最后，结合验证集表现和实际震例，展示并分析了三个震级预测模型的效果。

第六章，总结与展望，对全文所做工作和取得的效果进行了总结，并对后续工作进行了展望。

第二章　数据简介及数据预处理

第一章分析了地震监测和预测的相关背景和研究现状，本章将首先从 AETA 项目背景、系统架构、安装情况以及数据采集流程对 AETA 系统及其数据进行系统介绍，其次研究并实现 AETA 分层数据处理中的数据预处理层。在数据预处理层中，对设备停电导致的突跳数据进行自动化识别和清洗，对由于不可抗力因素导致的数据缺失进行分类补全，通过重新采样与频率转换解决了 AETA 系统升级导致的数据传输采样率不一致问题。

2.1　AETA 系统及其数据

2.1.1　AETA 系统简介

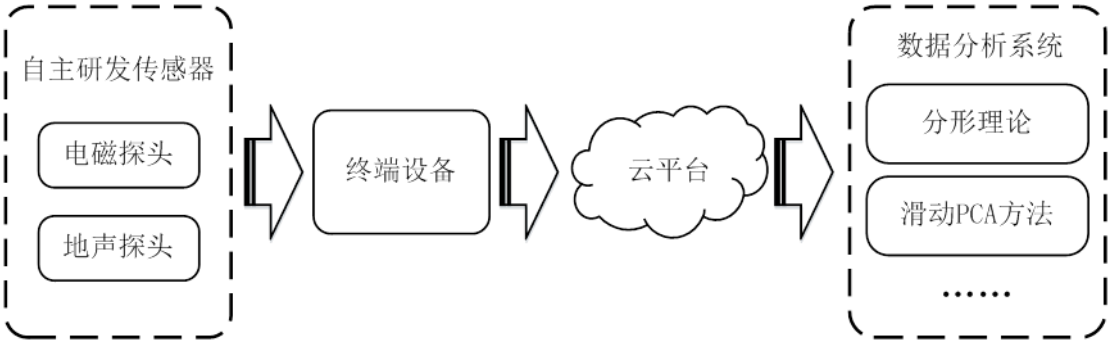


图 2.1　AETA 系统框图

AETA 系统由北京大学深圳研究生院集成微系统实验室研发，是一套可以大区域、高密度布设的观测地震前兆异常的软硬件综合体系[4]。该监测系统由地声探头、电磁探头、数据处理终端以及监测数据云平台和数据分析系统组成（图 2.1）[68]。地声探头监测 0.1Hz~50kHz，覆盖次声波、可听声波以及部分超声波波段，灵敏度为 3LSB/pa@0.1Hz~50kHz，18 位分辨率，低频采样率 500Hz，全频采样率 150kHz；电磁探头监测 0.1Hz~10kHz，覆盖 0.1~1000 nT 较宽动态范围的甚低频、超低频电磁波段，灵敏度>20 mV/nT@0.1 Hz~10 kHz，噪声水平为 0.1-0.2pT/Hz@(10Hz-1kHz) ，18 位分辨率，低频采样率 500Hz，全频采样率 30kHz[69]。该系统的目标是，通过密集布设的无人值守台站，感知来自地下的电磁扰动和地声信号，实时采集数据并通过互联网（有线或无线）将数据传输到云平台进行后续存储、处理，并基于采集数据研究地震活动、

服务于大地震的短临预测[68]。

AETA 系统于 2015 年 8 月完成了第一版设备的小批量试制，于 2016 年 6 月完成了第二版设备的小批量生产。在中国地震局监测预报司的支持以及河北、四川、云南、北京地震局协助配合下，两版设备均进行了现场布设试验。现场试验证明，AETA 系统对当地震例具有较好的映震效果，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深度合作，完成了第三版设备的二次开发和生产，AETA 系统正式步入定型批量生产阶段。截止目前为止，AETA 系统，在全国范围内安装超过 200 台，遍布河北、四川、云南、西藏、广东和台湾地区，具体分布见图 2.2。

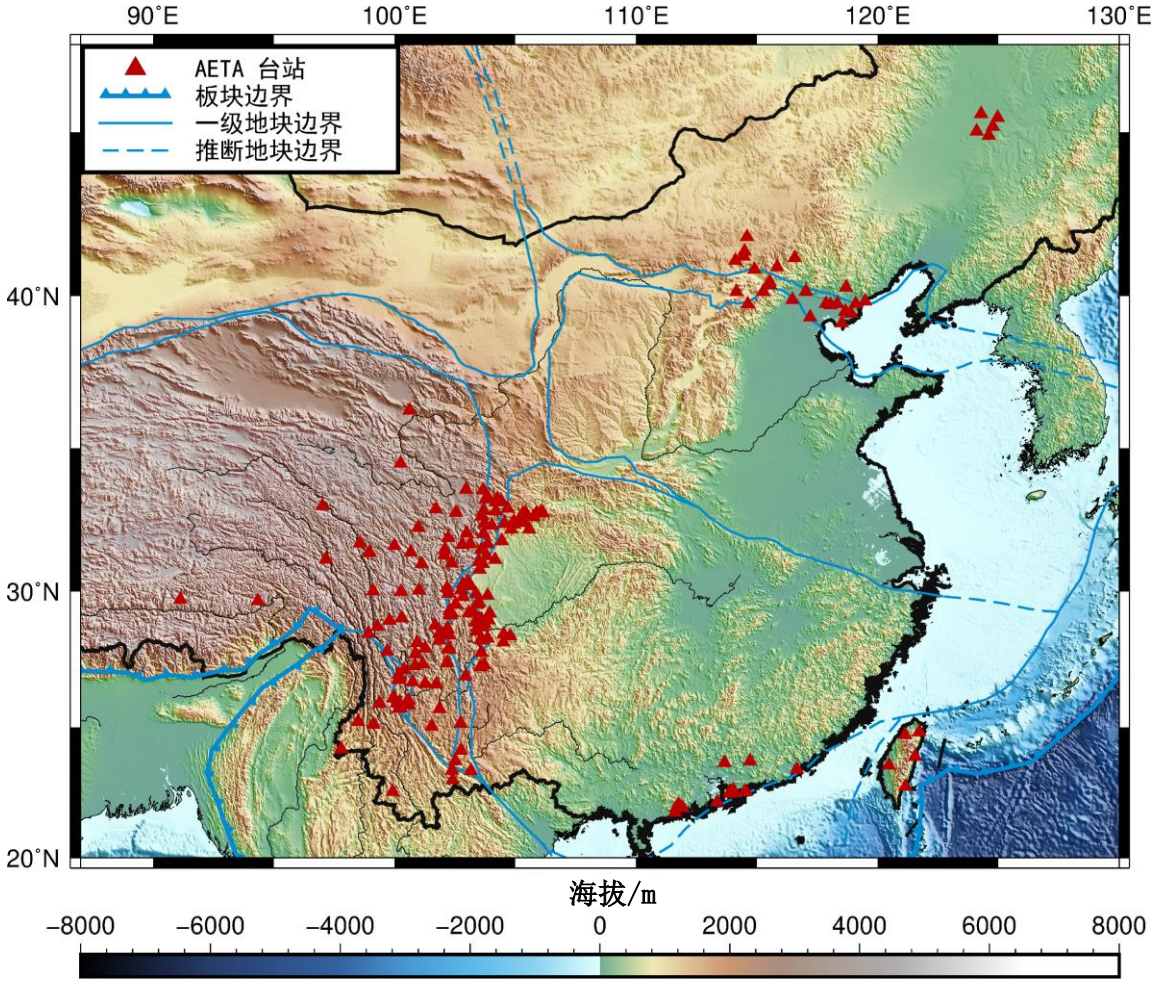


图 2.2　AETA 系统站点分布图

2.1.2　AETA 数据采集流程

AETA 数据由电磁、地声传感探头采集，在经过数据处理终端进行简单的抽样和滤波后，发送至应用服务器进行分类处理，并最终存储到数据中心，具体的数据采集

流程如图 2.3 所示。

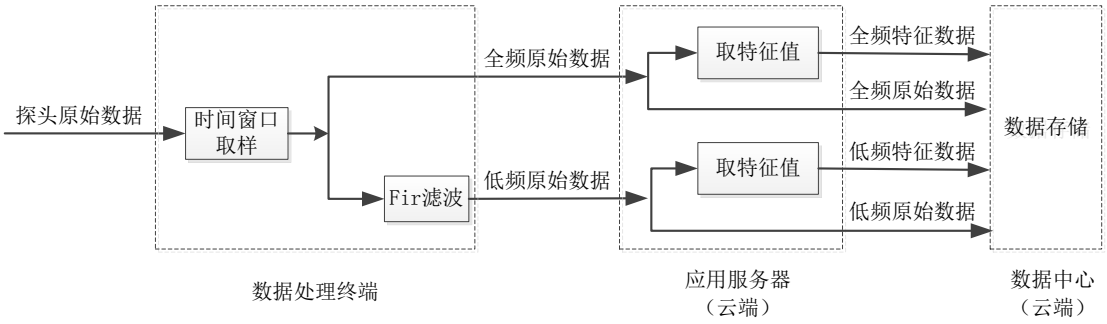


图 2.3　AETA 系统的数据采集流程

探头采集到地震前兆信号并转化为原始数据后，会通过 TCP/IP 协议把数据传输到数据采集终端。此时，地声数据的采样频率为 150kHz，电磁数据的采样频率为 30kHz。

在数据处理终端内中，会通过窗口取样的方式对数据量巨大的探头原始数据进行等间隔抽样，并把抽样的数据存入 Buffer。Buffer 内数据会分为两路，一路经软件低通滤波后成为低频原始数据（采样率 500Hz）；另一路不作任何处理，以全频原始数据的形式保留下来。最后，数据终端通过 HTTP 协议把原始数据发送至应用服务器。从 2017 年 6 月稳定运行至今，AETA 系统对数据处理终端进行了两次调整。第一次调整于 2017 年 8 月至 10 月进行，所有已安装台站的抽样策略陆续从初始版本的每 10 分钟抽取 1 分钟原始数据调整为每 3 分钟抽取 1 分钟原始数据；第二次调整发生在 2018 年 11 月至 2019 年 3 月，取消了全频原始数据，低频原始数据也从时间窗口取样变为连续采集。

应用服务器把原始数据分支为两路，一路不作处理并以原始数据的形式存入数据存储层；另一路从原始数据中提取出均值、振铃计数和峰值频率等基础特征值，以特征数据的形式存入数据存储层[69]。本文在第二至五章所开展的工作，都基于特征数据，下文所述地声数据、电磁数据皆指特征数据，表 2.1 是特征数据的基本信息。

表 2.1　AETA 系统中的特征数据

信号分量	频率范围	数据传输采样率
低频电磁	0.1Hz-200Hz	1 条/1min、1 条/3min、1 条/10min
全频电磁	0.1Hz-10kHz	1 条/1min、1 条/3min、1 条/10min
低频地声	0.5Hz-200Hz	1 条/1min、1 条/3min、1 条/10min
全频地声	0.5Hz-50kHz	1 条/1min、1 条/3min、1 条/10min

2.2 数据预处理

数据预处理是在进行数据挖掘等数据处理手段之前，对数据进行的一些前置处理。完整性和一致性都较差的脏数据和异构数据占据了实际数据中的绝大部分，若直接对其进行数据挖掘，效果往往不如人意。因此，致力于提升数据质量的数据预处理技术应运而生。数据预处理有多种方法：数据清理，数据填补，数据变换等。这些数据处理技术能大大提高数据质量，降低后续数据挖掘所需要的时间和难度。在工程实践中，数据预处理没有万能的方法，也没有标准的流程，通常需要根据任务和数据集的特性灵活处理。

结合 AETA 数据，本文在数据预处理阶段主要完成三方面的工作：(1) 断电数据处理，主要是清洗设备断电导致的地声突跳数据；(2) 缺失值处理，依据数据缺失的连续时长不同进行分类处理；(3) 重新采样与频率转换，兼容 AETA 数据中的多种数据采样率。

2.2.1 断电数据处理

在 AETA 系统中，部分台站的早期设备会在断电重启后出现的不合理地声跳变数据。这些由断电导致的跳变数据和台站接收到强烈地声信号所产生的数据较为相似，若不对其进行处理，会给后续的数据分析带来极大的困扰。在数据科学领域，过滤这些异常的不合理数据，亦即所谓的“脏数据”，其关键点在于如何定义和识别数据的合理性，常用的识别方法有以下三类：(1) 统计分析方法，如偏差分析、分布或回归方程识别非法点等；(2)构建规则库检查数据值，规则一般包括常识性规则、业务特定规则等；(3) 引入外部信息来检测和清理数据[70]。

结合 AETA 系统日志信息和地声跳变数据的模式，本文实现了一个自动识别断电跳变数据的算法，并对识别到的断电数据进行删除处理。自动识别算法首先通过数据库中的重启记录获得各个台站待筛选的时间点，其次结合台站失联时间以及失联前后信号的变化幅度锁定跳变数据的起始点。对于一段具体的跳变数据，先统计台站失联前后一天的数据，并根据数据分布中的四分位、方差等参数以及 3sigma 原则得出该台站数据的正常波动范围。最后，综合考虑异常数据长度、数据正常波动范围、趋势方向变换等因素，利用投票机制获取跳变数据的截止点。图 2.4 是三类常见的断电导致的地声跳变数据及其自动识别效果。红框部分为算法自动识别的地声跳变数据，可以看到该算法能识别“小 L 形”、“N 字形”、“长尾形”等不同形态的地声跳变数据。

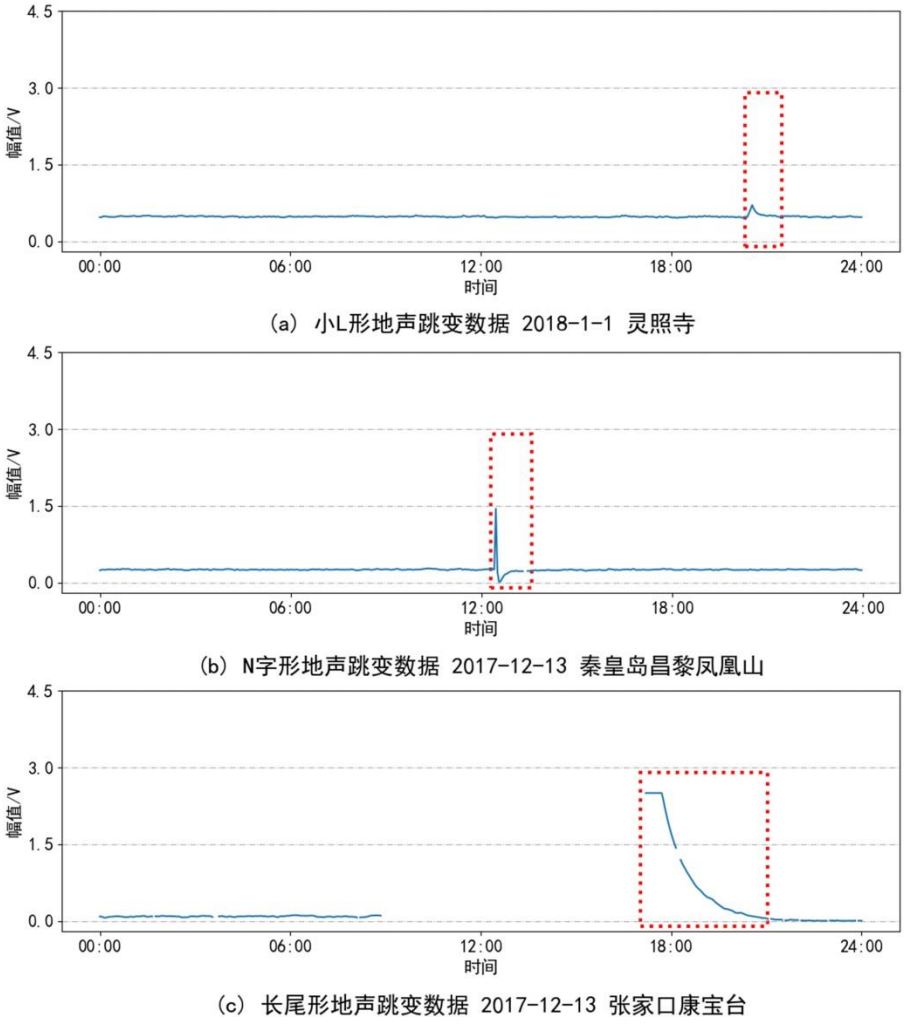


图 2.4　断电导致的地声跳变数据及其自动识别效果

2.2.2　缺失值处理

缺失值处理的目标是处理数据集中的缺失部分，从而避免数据缺失可能带来的一系列问题。例如，基于完整数据集的统计特征和分布特征会因数据缺失导致对于数据的估计从无偏变为有偏。以具体的均值和方差为例，缺失数据会改变数据的统计特征和分布特征，所得的具体数值就是带有偏差的结果。对缺失值进行处理，也契合了机器学习、数据挖掘等算法对数据集的要求，为下文建立地震风险模型提供了坚实的数据基础。

常用的缺失值处理方法有：基于数据统计的均值插补、多重插补、极大似然估计；基于模型的建模预测、高维映射等[71–73]。AETA 数据的缺失，是由系统故障、网络问题、断电数据处理等原因引起的，这属于时间序列数据缺失模式中的单变量数据缺失模式。对于这类缺失模式，一般采用简单的处理方法，而本文根据缺失数据的长度，

采用忽略缺失值和线性插补两种方式。

针对连续缺失少于 12 个小时的数据，本文采用线性插补处理缺失值。线性插补算法是插补中一种最常见的方法，主要思路是根据缺失数据的起始点和终结点进行线性插补。这种方法相比于均值插补、众数插补而言，更加注重时间序列的变化趋势，不会引入新的突变点，符合 AETA 数据对缺失值填补的要求。对于缺失超过 12 个小时的数据，为了不影响后续的数据处理，本文会忽略缺失值，不对数据进行填补。因此，AETA 的电磁和地声数据在经过数据预处理之后仍然会存在缺失值，而基于这些数据所提取的特征序列以及异常序列也会存在缺失值，这些处于第二、三层的缺失值，本文会针对性地使用均值填充、特殊值填充、多重填补等方法进行缺失值处理，这里不进行赘述。

2.2.3 重新采样与频率转换

重新采样是指对时间序列的采样频率进行转换的过程。重新采样可分为向上采样和向下采样两种。向上采样是指从低采样率转换到高采样率，此时时间序列的数据点会增多，通常使用插值的办法来增加数据点。向下采样是指将更高采样率的数据聚合到低采样率的数据中，此时时间序列的数据点会减少，往往采用箱体抽样的办法减少数据点。

对于时间序列这类按照时间排序的一组随机变量而言，采样率是一个重要的参数，它决定了对某种潜在过程进行观测的时间间隔。若某一个时间序列的采样率不能保持一致，则会对之后的数据处理带来困难。由于版本的升级变更，AETA 数据存在多个采样率，因此需要通过重新采用的办法来统一采样率。但不同的数据处理方法，对于重新采样的要求也各不相同。例如本文第三章所提及的分形维数计算方法，关注的是数据的混沌程度（亦可简单理解为数据曲线的不规则性），此时若使用向下采样，会丢失大量的数据细节，降低了分形维数这一特征值的数据有效性，因此应使用向上采样来统一采样率。而对于像均值、最大值等统计特征，向上采样所产生的数据冗余对后续处理没有任何帮助，所以应采用向下采样来统一采样率。

为满足 AETA 数据统一采样率的需求，本文实现了一个重新采样模块。该模块可将 AETA 数据按照指定的采样率进行重新采样，支持的采样率为 1 条/1min、1 条/3min、1 条/10min，具体流程如图 2.5 所示。

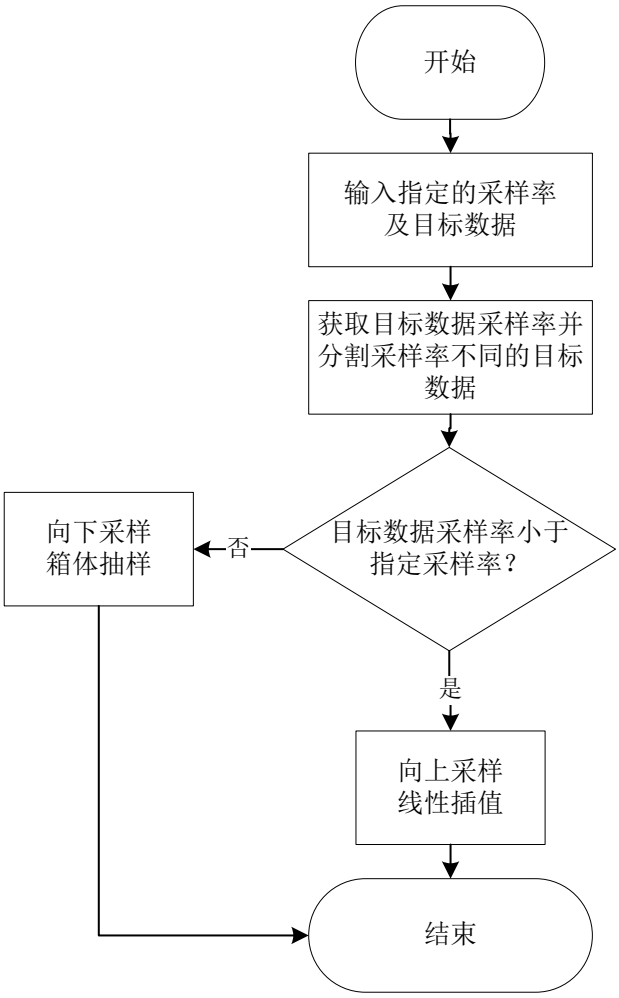


图 2.5　重新采样模块程序框图

2.3　本章小结

本章首先系统介绍了 AETA 的项目背景、系统架构、观测探头、安装情况，并从数据采集流程的角度对 AETA 数据进行全面梳理。其次，实现 AETA 分层数据处理中的数据预处理层。在数据预处理层中，本文主要完成以下三项工作：1. 对设备停电导致的突跳数据进行自动化识别和清洗，所实现的自动识别算法能区分正常地声变化和停电导致的地声跳变数据，并能识别多种形态的地声跳变数据。2. 对由于不可抗力因素导致的数据缺失，根据缺失数据的长度分别用忽略缺失值和回归插补的方式进行处理。3. 实现了一个可将 AETA 数据按照指定的采样率进行采样的重新采样模块，解决了 AETA 系统升级导致的数据传输采样率不一致问题。经过数据预处理后，AETA 数据的数据质量得到提高，为下文实现其他数据处理方法提供了良好的数据基础。

第三章　特征提取与异常检测

第二章已对 AETA 数据进行了预处理，本章节从特征提取和异常检测这两个层次对 AETA 数据进行处理。通过特征提取挖掘 AETA 数据中的丰富内涵，通过异常检测挖掘特征序列的变化趋势，并最终生特征序列和异常指数序列。

3.1　特征提取

基于时间序列的分类任务中，提高分类准确率途径有两种： 一是改进分类器； 二是采用特征提取技术。特征提取，是指在分类器进行数据分类之前，通过对数据进行合理的归约以减少数据冗余，从而降低分类难度并提高分类准确率。

一般而言，时间序列都具有趋势性、周期性，而不同领域的时序数据又带有一些领域特征。如金融数据，自相关属性明显，且在趋势上往往呈现“高峰厚尾”的特点；语音信号，信号的幅值主要分布在零附近，这是由于语音信号不持续的特性导致的；心电信号，以周期性著称，频域重复率低且各波段的频率相对独立[74]。时序数据的特殊性，也对特征提取提出了更高的要求，特征提取所得的特征矢量既要保持原有时间序列的性质，又要凸显不同数据的区别，这样才能保证最终的分类效果。针对不同的数据，常见的特征提取方法如表 3.1 所示。

表 3.1　常见的特征提取方法

类别		特征矢量	体征提取的主要方法
基本统计方法		基本统计特征作为特征矢量	均值，方差，极值， 波段，功率谱，过零点等
基于分形理论		分形维数作为特征矢量	相似维数、盒维数、关系维等
基于模型		模型的系数作为特征矢量	自回归滑动平均模型、滑动平均模型、组合-ARMA 模型、自回归条件异方差族计量模型等
基于变换	频频	利用变换后的频域系数等作为特征矢量	傅立叶变换、倒谱系数
	线性	主成分或小波系数等作为特征矢量	PCA、K-L 变换、小波变换

AETA 数据作为一种传感器监测数据，是典型的时序数据。时序数据与时间相关联，数据量大且单个数据点的信息含量不高，若直接使用机器学习方法处理 AETA 数据，很可能会陷入训练困难、无法拟合的困难，因此需要对 AETA 数据进行特征提取。本节将研究并实现基于基本统计方法提取 AETA 数据的统计特征和基于分形理论提取 AETA 电磁数据的分形维数。通过特征提取，可以获得多种采样率为一天一个点的特征序列，并可作为异常检测、地震风险模型的输入。

3.1.1　基于统计方法的特征提取

常见的基本统计特征可分为两类：(1) 时域包括均值、方差、极值、过零点、最值等；(2) 频域包括峰值频率、功率谱密度、相频特性、平均功率频率等。本文针对 AETA 数据的时域、频域特征，特别是对 SRSS 波的理解，提出了 13 种统计特征。前 6 个特征是在时域上 AETA 电磁、地声数据的基本统计值；第 7~9 个特征是为了刻画某些震例中地声数据的震前异动；第 10~13 个特征则是根据 SRSS 波的日升日落特性，泛化出来的对电磁上升沿及下降沿的刻画，具体如表 3.2 所示。

表 3.2　AETA 数据的统计特征

特征类型	特征名称	缩写	特征意义	理论取值范围
基本统计特征	电磁均值	EMA	电磁数据一天的均值	0~12.288
	地声均值	GSA	地声数据一天的均值	0~2.5
	电磁方差	EMV	电磁数据一天的方差	0~+∞
	地声方差	GSV	地声数据一天的方差	0~+∞
	电磁最大值	EMM	电磁数据一天中的最大值	0~12.288
	地声最大值	GSM	地声数据一天中的最大值	0~2.5
地声异动特征	尖峰个数	Peak	地声数据在一天中出现的尖峰个数	0~144
	翻转次数	Sway	地声数据的一阶导数的变化次数	0~144
	峰值频率	PF	地声一天的峰值频率数据中，除去50Hz，占比最高的频率	0.1~10000
SRSS 波特征	最大上升值	RM	电磁数据一天中的最大差分值	0~12.288
	最大上升时间点	RMP	最大上升值对应的数据点序号	0~1440
	最大下降值	DM	电磁数据一天中的最小差分值	-12.288~0
	最大下降时间点	DMP	最大下降值对应的数据点序号	0~1440

3.1.2 基于分形理论的分形维数提取

与微积分所描述的平滑变化不同，分形是指在无限放大下仍然具有不规则突变的具有无穷自相似结构的点集，是讨论不规则的一套语言，讨论的是局部与整体的自相似性，大自然中的海岸线、雪花都可分属于分形[75]。虽然只经历了十余年的发展，但分形学已经成为非线性科学的两大重要组成部分之一，在特征提取的应用超凡出新，在时间序列中的应用更是硕果累累。如在机器故障的诊断中，常把机器运作信号划分为多种状态，此时可针对不同状态提取分形维数，并分析分形维数与机器故障的相关性，以此达到故障诊断的效果。

分形理论用于特征提取时，主要是计算非线性信号的分形维数。作为一种定量的分析指标，分形维数可以描述研究对象的混沌程度和表征空间的扩展程度。根据定义不同，分形维数包括相似维数、豪斯道夫维数、盒维数、空间维数等。在实际应用中，分形维数被用来描述旱洪灾的内在变化规律[76]、检测股票离群点[77]、对金融管理或风险数据进行分类[78]。

一维时间序列分形维数的计算方法可分为两种[79]：(1) 重构时间序列的相空间并在相空间中计算关联维数[80,81]；(2) 在时间域内计算分形维数。秦建强等在 2016 年以 WCF 合成时间序列为研究对象，分别对 Katz 方法、Castiglioni 方法、Sevcik 方法、盒维数方法、基于 FA 的计算方法、基于 DFA 的计算方法、基于 R/S 分析的计算方法和 Higuchi 方法这八种常用的分形维数算法的准确性和效率进行分析对比，实验结果表明 Higuchi 算法在计算一维时间序列的分形维数时在准确性、效率上都有不俗的表现[82]。

AETA 电磁扰动作为一种复杂信号，本文通过计算其全天数据的分形维数定量地描述电磁扰动当天的混沌程度，以下是具体的计算方法：

设电磁扰动的全天数据为$\{x_n\}$，序列长度为 N。根据抽样间隔 k，抽样得到 k 个长度为$\mathrm{int}(\frac{N}{k})$的抽样序列，其中第 m 个抽样序列：

$$x_k^m=\left\{x_m,x_{m+k},x_{m+2k},\cdots,x_{m+\mathrm{int}\left(\frac{N-m}{k}\right)\cdot k}\right\},\quad m=1,2,3,\cdots,k \tag{3.1}$$

计算这 k 个抽样序列的曲线长度$L_m(k)$以及它们的平均曲线长度$L(k)$：

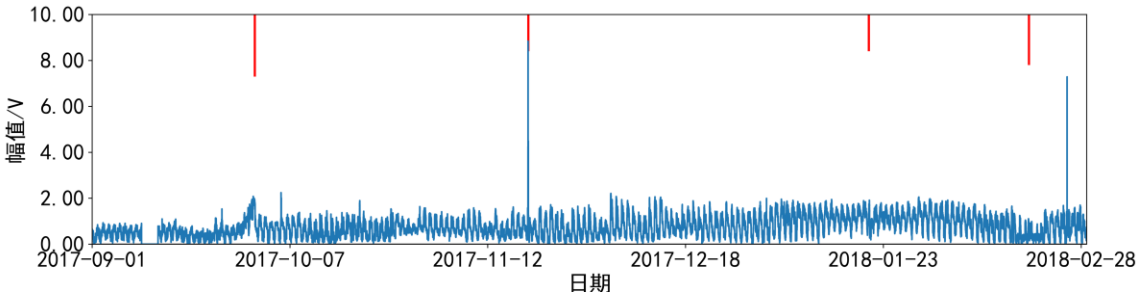
$$L_m(k)=\frac{N-1}{\mathrm{int}\left(\frac{N-m}{k}\right)\cdot k^2}\cdot\sum_{i=1}^{\mathrm{int}\left(\frac{N-m}{k}\right)}\left|x_{m+ik}-x_{m+(i-1)k}\right| \tag{3.2}$$

$$L(k)=\frac{1}{k}\cdot\sum_{m=1}^{k}L_m(k) \tag{3.3}$$

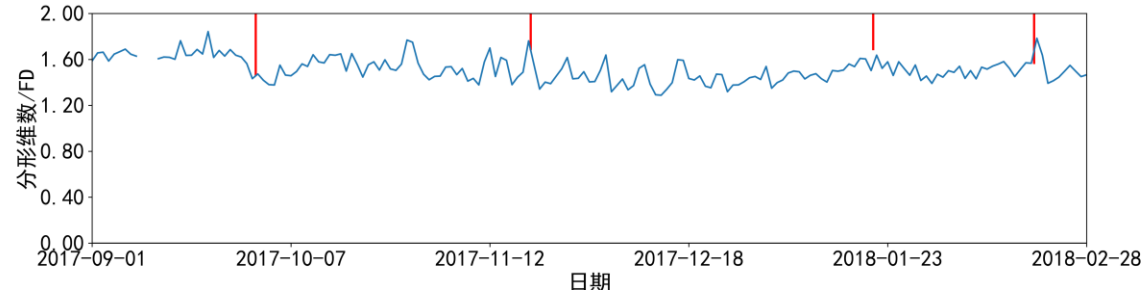
从而得到抽样间隔 k 与平均曲线长度L(k)的关系。对于一个台站一天的电磁扰动特征数据$\{x_n\}$，如果该序列可以分形，则序列的所有$(k, L(k))$对满足以下条件：

$$L(k)\propto\left(\frac{1}{k}\right)^{FD},\quad k=1,2,3,\cdots,k_{max} \tag{3.4}$$

其中 FD 就是电磁扰动全天数据的分形维数，为了提高 FD 的准确率，本文将k_{max}设置为$\frac{N}{2}$。最后将k_{max}个$\left(\frac{1}{k}, L(k)\right)$数据点放在双对数坐标上，通过最小二乘法拟合出一条直线，直线的斜率就是电磁扰动全天数据$\{x_n\}$的分形维数 FD。FD 描述序列$\{x_n\}$的混沌程度，数值越大混沌程度越高。利用 HFD 算法对多天的电磁扰动数据进行计算，可以得到一个分形维数序列{FD}，该序列描述电磁扰动混沌程度的变化趋势，其中FD_i代表第i天电磁扰动的混动程度。



（a）　平武县防震减灾局台站的电磁扰动



（b）　平武县防震减灾局台站的分形维数

图 3.1　平武县防震减灾局台站的电磁扰动数据、分形维数

以平武县防震减灾局台站从 2017 年 9 月 1 日到 2018 年 2 月 28 日采集的电磁扰动数据为例，图 3.1 (a)为该台站的电磁扰动数据，图 3.1 (b)为使用 Higuchi 方法得到的电

磁扰动的分形维数的天变化波形。为了更好地展示异常的映震效果，将震中距小于55km 震级大于 2.0Ms、震中距小于 110km 震级大于 4.0Ms 的地震事件以红线的形式标注在图 3.1 中，震级越大线段越长。图 3.1 所示的四个地震分别是 2017-09-30 青川 Ms5.4 地震、2017-11-19 平武 Ms3.2 地震、2018-01-20 北川 Ms3.2 地震以及 2018-02-18 青川 Ms4.4 地震。

3.2 异常检测

上一小节所提取的特征值从一定程度上可以反映地震活动，但相比于特征值的大小，特征值的异常变化往往更能刻画地震活动。同时，由于安装地点、地下环境的不同，不同台站在某些特征值上的差距会比较巨大。

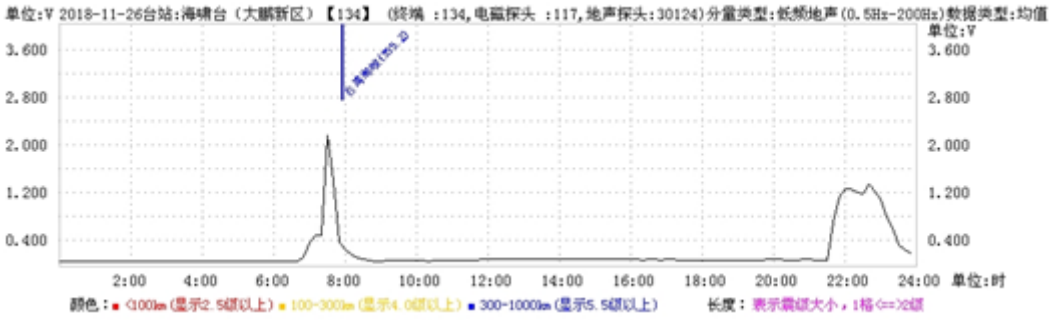


图 3.2 大鹏新区海啸台地声均值的地震前兆

例如，大鹏新区海啸台的地声均值长期保持在 0.1 伏左右，青川乔楼的地声均值长期保持在 2.0 伏左右。因此，0.1 伏、2.0 伏都应该是正常的地声均值，不应被归为地震前兆。而在 2018 年 11 月 26 日台湾海峡发生 5.2 级地震时，大鹏新区海啸台的地声均值发生突变，如图 3.2 所示，在震前两小时内地声均值从 0.1 伏上升到 2.1 伏左右，此时的 2.1 伏是一个明显的地震前兆，但若仅通过地声均值的大小来判别地震活动的话，很可能会出现错判的现象。因此，本文需要通过异常检测的方法来处理 3.1 小节所得到的特征数据。这样一来保留时序数据的时序特点，二来从某种程度上是对不同台站进行归一化。

异常检测始于 20 世纪 80 年代，具体可分为以下几类：

1. 基于统计学方法。检验假设是最早的统计学异常检测方法，该方法首先对数据集的分布模型作出假设，其次基于假设分布把小概率事件检测为异常。在实际使用时，数据集的分布往往难以预知，所提出的分布模型与数据的契合程度有限，导致检测效果一般。

2. 基于距离、密度的方法。这类方法通过观察正常数据与异常数据在数据对象最

近邻距离、数据对象所在区域密度的不同，选择合理的阈值进行异常判断。此类方法在由不同密度的数据混合而成的数据集上效果一般，而且由于使用硬阈值，泛化能力也不强。

3. 基于聚类的方法。依据所擅长的数据集不同，聚类方法可大致分为三类：第一类，适用于正常数据实例处于某些集群中而异常数据远离所有集群的情况。这些聚类方法不强制每个数据实例归属于具体的集群，例如 DBSCAN、普聚类等聚类方法；第二类，适用于正常数据靠近它们聚类中心而异常数据则远离聚类中心的情况，常见方法为 SOM(Self-Organizing Maps)、K-means 等；第三类，适用于正常数据的所在集群，其规模和密度都比异常集群大的情况，find_CBLOF 算法是个中翘楚。聚类方法众多，在选取具体算法时，需要根据数据的特征进行选择。

4. 基于时序数据特性的方法。此类方法从时序特征入手进行异常检测，可分为以下四种：第一种，基于特征空间的片段分割把数据投射到特征空间，常用的是 PRL 表示法；第二种，基于已有数据建立回归模型，常用的有 AR、ARMA、ARIMA 等；第三种，基于数据频率的编码方法，例如马尔科夫模型、后缀树检测异常模式[83]；第四种，基于子序列特征的相似性判断，以 SAX 为主要代表。基于时序的方法，大多需要数据在时序上具有一定的模式，在模式不明显的数据集上表现一般。

综合上面的分析可以知道，虽然时间序列的异常检测方法众多，但是不同方法都有各自的局限性，在实际应用中并没有标准化的检测方法。本节在 AETA 数据上将实现两种不同的异常检测方法。对于一维数据，本节使用滑动四分位距法计算异常值；对于多维数据，本节使用 DBSCAN 密度聚类进行异常点检测。通过异常检测，可以获得多种采样率为一天一个点的异常序列，并供下文所述的地震风险模型使用。

3.2.1　基于滑动四分位距法的异常值计算

四分位距（IQR）是一种用来表示稳健统计中数据离散度的量，可用来检测数据的异常情况[84]。滑动四分位距法是对数列进行滑窗取值后，利用滑窗内 IQR 对各个数值进行异常检测并生成异常值的方法。张明敏等使用滑动四分位距法法研究九寨沟 7.0 级地震前电离层 TEC 的异常取得了一定的效果[85]。本文使用滑动四分位距法对 AETA 数据的分形维数特征进行异常值计算。

假设单个台站的分形维数序列{FD}在一个时间窗口内服从稳定分布，若在此期间某一天的分形维数偏离周期内的稳定分布，则认为该天出现电磁扰动异常变化。通过滑动四分位距法标记并量化这种异常，具体如下：

设分形维数序列为$\{\mathrm{FD_n}|\mathrm{n}=1,2,3,\cdots,\mathrm{N}\}$，异常值序列为$\{\mathrm{Outlier_n}|\mathrm{n}=\mathrm{L},\mathrm{L}+1,\cdots,\mathrm{N}\}$，滑动窗口宽度$\mathrm{L}=27$。其中，异常值序列是分形维数序列经滑动四分位距法

处理所得，窗口宽度的选取则是根据太阳辐射周期为 27 天，选取该窗口长度可排除太阳活动对电磁扰动的干扰[86]。

异常值序列中异常值$\mathrm{Outlier_n}$由待检测值$\mathrm{FD_n}$与对应的窗口ω_n的异常值探测区间共同决定。$\mathrm{Outlier_n}$所对应的窗口ω_n为：

$$\omega_n = \{\mathrm{FD_i} | i = n - L + 1, n - L + 2, n - L + 3, \cdots, n\} \tag{3.5}$$

窗口ω_n的 IQR 值及相应的异常值探测区间为：

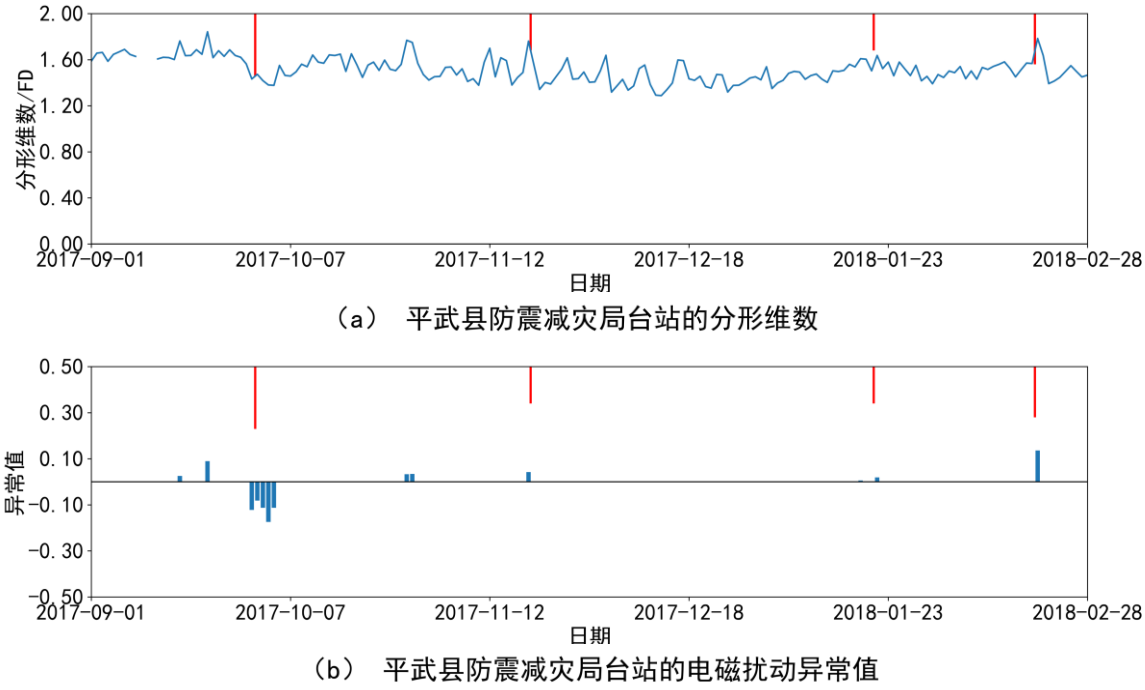
$$\begin{gathered} \mathrm{IQR} = \mathrm{Q1} - \mathrm{Q3} \\ [\mathrm{Q2} - \mathrm{IQR} \cdot \mathrm{k}, \mathrm{Q2} + \mathrm{IQR} \cdot \mathrm{k}] \end{gathered} \tag{3.6}$$

Q1、Q2、Q3分别为第一、第二、第三四分位数。根据标准化正态分布下四分位距（IQR）与标准差的关系 IQR=1.34 σ，本文中参数k选取了典型值 1.5。

若待检测值$\mathrm{FD_n}$在异常值探测区间内，待检测值服从稳定分布，异常值$\mathrm{Outlier_n} = 0$；若待检测值$\mathrm{FD_n}$在异常值探测区间外，待检测值$\mathrm{FD_n}$为异常值（置信度 99%），异常值$\mathrm{Outlier_n}$的具体计算如下：

$$\mathrm{Outlier_n} = \begin{cases} (\mathrm{Q2} - \mathrm{IQR} \cdot \mathrm{k}) - \mathrm{FD_n} & x_i < (\mathrm{Q2} - \mathrm{IQR} \cdot \mathrm{k}) \\ 0 & x_i \in [\mathrm{Q2} - \mathrm{IQR} \cdot \mathrm{k}, \mathrm{Q2} + \mathrm{IQR} \cdot \mathrm{k}] \\ \mathrm{FD_n} - (\mathrm{Q2} + \mathrm{IQR} \cdot \mathrm{k}) & x_i > (\mathrm{Q2} + \mathrm{IQR} \cdot \mathrm{k}) \end{cases} \tag{3.7}$$

异常值$\mathrm{Outlier_n}$定量地描述台站当天电磁扰动混沌程度的异常情况，数值越大异常程度越高。同样地，取平武县防震减灾局台站从 2017 年 9 月 1 日到 2018 年 2 月 28 日的数据为例，如图 3.3 所示，图 3.3 (a)为电磁扰动的分形维数的天变化波形，图 3.3(b)是使用滑动四分位距法提取图的异常值。相比图 3.3(a)，图 3.3(b)的异常值数据与地震活动的关系更为直观明显。



（a）　平武县防震减灾局台站的分形维数

（b）　平武县防震减灾局台站的电磁扰动异常值

图 3.3　平武县防震减灾局台站的分形维数和电磁扰动异常值

为进一步论证方法的有效性，本文选取低阻—中低阻的龙门山断裂带中北段区域作为实验范围。基于本文所提出的异常提取方法，对实验范围内九个 AETA 台站的电磁扰动数据进行实验，台站位置如图 3.4 所示，台站的地理位置信息见表 3.3。

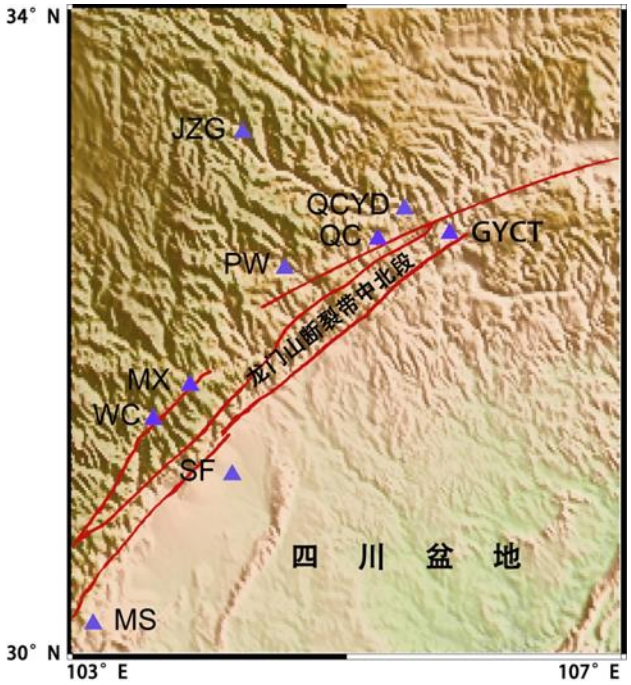


图 3.4　研究台站分布图

表 3.3　台站位置信息

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

除了图 3.3 所示的平武县防震减灾局台外，其余 8 个台站相同时期内电磁扰动分形维数的异常值如图 3.5 所示。部分地震学家认为，电磁扰动异常或许是孕震期区域构造应力变化所引起的机电转换机制导致的，这意味着电磁干扰异常可能在孕震期继续发生，所以本章把同一台站间隔小于等于 5 天的多个异常（异常指数不等于 0）合并为同一个异常。以平武县防震减灾局台站为例，9 月 16 日和 9 月 21 日出现的异常合并为一次异常，9 月 29 日到 10 月 3 日连续五天出现的异常合并为另一次异常。图 4 和图 5 表明，在 2017 年 9 月 1 日到 2018 年 2 月 28 日半年间，在 9 个研究台站附近发生了 27 次地震，其中 23 次地震发震前 15 天到震后 5 天内至少有一个周边台站出现异常，占地震总数的 85.2%。结果表明，该方法提取的电磁扰动异常，有一定的映震效果。

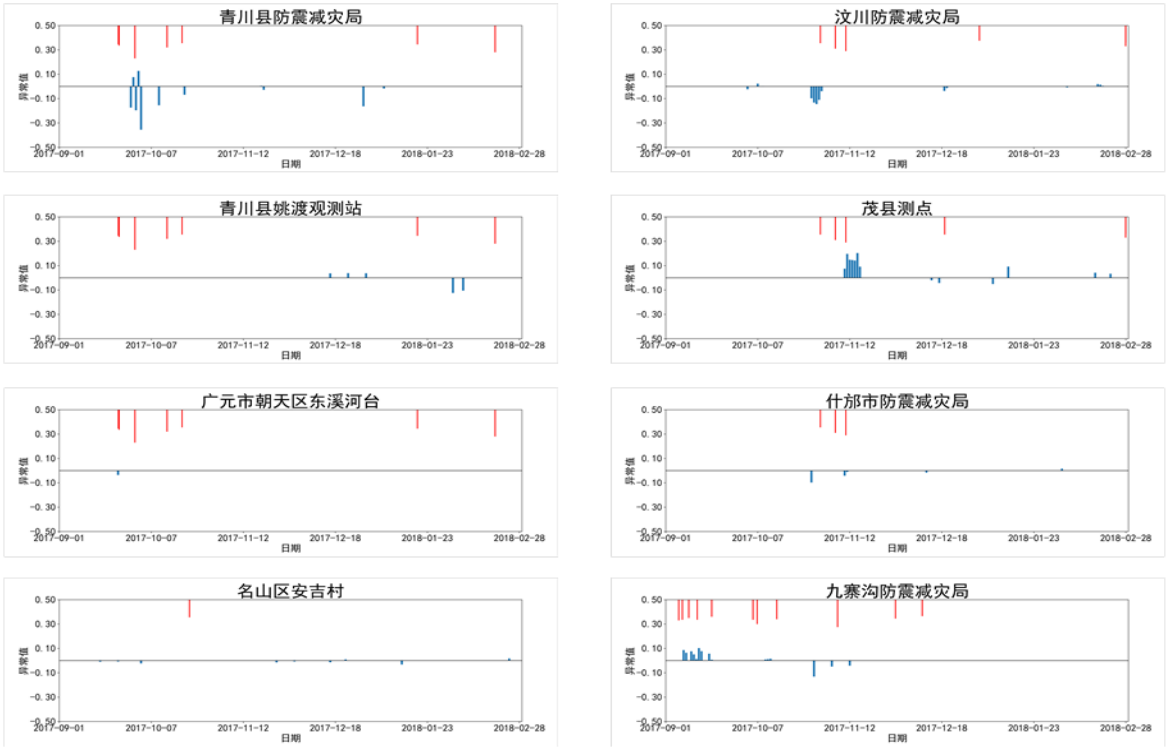


图 3.5　各研究台站电磁扰动异常值与附近地震

3.2.2　基于 DBSCAN 密度聚类的异常点检测

聚类是通过学习数据内在分布的规律，找出数据簇的方法。对于无监督的聚类任务，聚类效果可以通过考察其聚类结果的“簇内相似度”和“簇间相似度”来判断。常用的评价指标有 DB 指数和 Dunn 指数。根据聚类原理的不同，常用的聚类方法可以分为原型聚类、层次聚类和密度聚类。

1. 以 K-means 和 LVQ 为代表的原型聚类，是通过对原型进行初始化后进行迭代求解的方法。此类方法不适合应用于 AETA 的数据，原因如下：(1) 原型数量 K 作为超参数，需要预先设定，且簇数的大小会直接影响聚类效果。而对于 AETA 数据，簇数并不是一个固定的值，以台站的电磁数据为例，簇数的典型值为 1~3。本文关注的是 AETA 数据的异常检测，并不关心数据簇的个数。(2) 异常点会严重干扰原型聚类的结果。AETA 数据中除了与地震活动相关的异常点外，也会因台站周边环境的变化产生少量异常点，这些点的存在会严重影响原型聚类的效果。

2. 以 AGNES 为代表的层次聚类，试图在不同层次将数据分类，形成树型的数据结构，并采用自下而上的聚合策略或自上而下的分拆策略对数据进行聚类。但此类方法在具体实现时需要灵活控制不同层次的聚类粒度，而 AETA 数据的层次分类特性并不明显，因此此类方法并不适用于 AETA 数据。

3. 以 DBSCAN、SNN 为代表的密度聚类，主要依据数据的密度分布来进行聚类。

密度聚类不需要预先设定类别数量，且聚类结果没有偏倚，初始值对聚类结果影响不大。与 K-means 只适用于凸数据集不同，密度聚类方法可以适用于任意形状的数据集。当然，密度聚类相对于另外两类聚类方法，时间复杂度会更高，但本文基于聚类所进行的异常检测场景中样本数一般为 27，最多不超过 648（27 天*24 小时），对时间复杂度的要求并不高。

综上，本文将 DBSCAN 密度聚类作为 AETA 数据分层处理框架中的一种异常点检测方法。DBSCAN 密度聚类算法是一种基于样本分布紧密程度进行聚类的算法。DBSCAN 通过邻域距离阈值 ϵ 和邻域中样本个数阈值 MinPts 来描述邻域的样本分布紧密程度。基于这两个参数，DBSCAN 提出了 ϵ-邻域、核心对象、密度直达点、密度可达等概念。所谓的核心对象是指其 ϵ-邻域内至少包含 MinPts 个样本的对象；如果 x_i 位于核心对象 x_j 的 ϵ-邻域中，则称 x_i 由 x_j 密度直达；密度可达是指两点之间满足传递性密度直达。DBSCAN 的聚类思想是通过密度可达找到最大的密度相连集合，并将其作为最终聚类中的一个簇。具体算法如下：

表 3.4　DBSCAN 密度聚类算法

算法：DBSCAN 密度聚类算法
输入： $D=\{x_1,x_2,...,x_m\}$：样本集 ϵ：领域距离阈值 MinPts：领域汇中样本个数阈值
输出： C：簇聚类结果
方法： 步骤 1)初始化核心对象集合 Ω、聚类簇数 k、未访问样本集 Γ、簇聚类结果 C； 步骤 2)计算各样本间的距离，并寻找出所有的核心对象，获得集合 Ω； 步骤 3)在核心对象集合 Ω 中，随机抽取一个核心对象 O。更新当前类别序号 k=k+1，设当前簇核心对象队列 $\Omega_{cur}=\{O\}$、当前簇样本集合 $C_k=\{O\}$，更新 $\Gamma=\Gamma-\{O\}$； 步骤 4)从 Ωcur 中取出随机取出核心对象 O’，找出其 ϵ-邻域内的子样本集 $N\epsilon(O')$，其中可以归为当前簇的样本为 $\Delta=N\epsilon(O')\cap\Gamma$，更新 $C_k=C_k\cup\Delta$、$\Gamma=\Gamma-\Delta$、Ωcur=Ωcur$\cup(\Delta\cap\Omega)-O'$； 步骤 5) 当 Ω_{cur} 为空时，当前类簇的最大密度相连集合搜索完成，C_k 生成完毕，簇聚类更新为 $C=C\cup\{C_k\}$，Ω 更新为 $\Omega-C_k$；否则，进入步骤 4； 步骤 6)当 Ω 为空时，算法结束；若 Ω 非空，转到步骤 3。

以 AETA 数据统计特征的最大上升时间点 RMP、最大下降时间点 DMP 所构成的

特征空间为例，说明使用 DBSCAN 密度聚类对 AETA 数据进行异常点检测的具体过程和检测效果。对单个台站而言，每天可以计算出一对（RMP，DMP）值，排除外界影响，在一个时间窗口内台站的（RMP，DMP）应具有相似性 ，若在此期间有部分（RMP，DMP）发生偏离，可通过 DBSCAN 密度聚类将偏离点识别出来。

使用 DBSCAN 密度聚类方法处理 2017 年 8 月 8 日九寨沟地震前九寨沟防震减灾局台站的（RMP，DMP）数据。在该实验中，根据太阳辐射周期把窗口设为 27 天，对当天和前 26 天组成的数据进行聚类，若当天的数据所在簇的样本数小于 7（经验值），则当天会被识别为异常点。聚类的距离度量使用欧氏距离，邻域参数 ϵ 取样本点邻接距离的均值，MinPts 设置为 2。根据算法的聚类结果，可以判断每一天是正常还是异常的，对一个台站的数据进行滑窗处理，从而得到一个表示当天是否异常的二值时间序列。结果如图 3.5 所示，8 月 6 日到 8 月 8 日的（RMP，DMP）数据被识别为异常点。相比于 RMP、DMP，通过 DBSCAN 密度聚类得出的二值时间序列能更直观地表现地震活动。

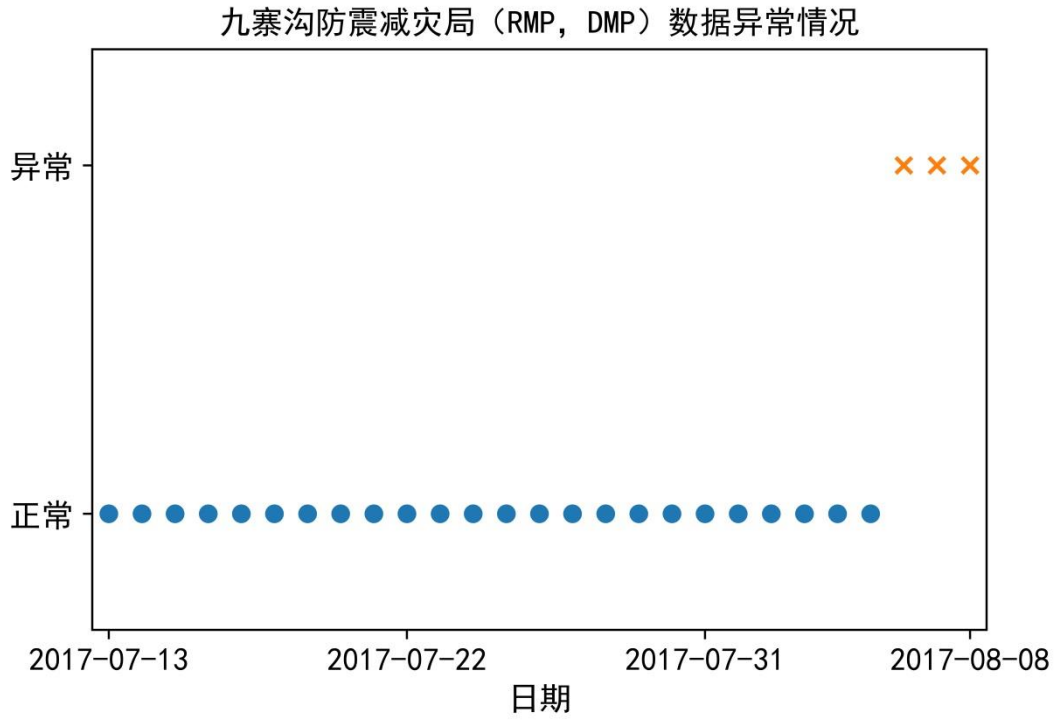


图 3.6　九寨沟防震减灾局震前 27 天的异常识别效果

3.3　本章小结

本章主要实现了 AETA 数据分层处理中的特征提取层和异常检测层。

针对特征提取，本章首先分析了 AETA 数据作为一种传感器监测数据的特点，并说明进行特征提取的必要性：数据信息密度低，容易导致模型无法收敛，因此无法直接输入到机器学习模型中。其次研究了特征提取的常用方法，并基于统计方法和分形理论对 AETA 数据进行了特征提取。基于统计方法提出了三大类统计特征：基本统计值、刻画地声震前异动、刻画 SRSS 波的日升日落特性，共 13 个统计特征；基于分形理论提取了 AETA 电磁数据的分形维数这一特征。

针对异常检测，本章首先结合大鹏新区海啸台以及青川乔楼的数据，分析了异常检测的重要作用：一是保留 AETA 数据的时序特点，二是从某种程度上对不同台站进行了归一化。其次研究了异常检测的常用方法，并依据 AETA 数据的特点，针对一维数据和多维数据分别采用滑动四分位距法和 DBSCAN 密度聚类进行异常检测。通过对 AETA 数据进行相关实验，说明经过异常检测后的数据更能反映地震活动。

事实上，在特征提取和异常检测这两个层次，还可以生成更加丰富的特征序列和异常序列，从更多维度描述 AETA 数据与地震活动的关系。例如，项目组其他成员所提出的基于滑动 PCA 方法生成的条带特征、基于希尔伯特-黄变换的频域特征等等。限于篇幅，本章只对本人在特征提取、异常检测所完成的工作进行说明。

第四章　地震风险模型的建立

地震预测中有一个关键的难题：如何把地震的三要素和监测信号关联，并从中寻找二者间的关系。本章基于地震三要素生成了风险标签，通过建立一个地震风险模型从上一章生成的特征序列和异常序列中寻找地震风险与监测信号的关系。该模型可以根据各个台站的数据，判断台站附近在近期是否有地震发生的风险。本章首先构建了地震风险模型的数据集，包括生成风险标签以及构建特征工程两方面。其次对分类算法进行选取，分析并对比所采用的随机森林算法和其他算法。最后，完成模型训练并展示模型的效果。

4.1　数据集构建

地震风险模型的目标是学习到地震三要素和监测信号的关系，这是典型的有监督学习任务。有监督学习指的是，从标签化的训练数据集中推断出输入到输出的映射函数的机器学习任务。在大多数的有监督学习任务中，标签都是明确的，如 MNIST 手写体识别中标签是数字 0 到 9、语音识别中标签是音频所对应的具体文字、物品分类中标签是苹果、雪梨等确定性的物体。但对于地震风险，并没有一个直接明确的标签可供使用。本节首先基于地震三要素生成地震风险模型中的样本标签，其次基于特征序列、异常序列等数据构建特征空间。

4.1.1　基于地震三要素生成样本标签

构建样本标签，本质上是确立学习任务所要求解的问题。样本标签是机器学习算法从特征空间中收敛的目标。地震风险模型的目标是通过台站的监测信号，判断在近期内台站附近有无发震风险，可抽象为一个二分类问题，正样本表示近期台站附近有发震风险，负样本表示近期台站附近无发震风险。本文认为，震级越大，影响的范围越大，前兆信号持续的时间越长。对于同一个地震事件而言，距离近的台站能判别为发震风险的天数更多，距离远的台站天数更少。为简化学习任务，降低模型的拟合难度，本文把地震事件按照震级分为弱震（Ms0~3）、有感地震（Ms3~4.5）、中强震（Ms4.5~6）、强震（Ms6 以上）。基于以上原则，本文从地震事件出发，根据台站的震中距分别把台站震前连续的 3、7、15 天标为正样本，即近期台站附近具有发震风险。震前连续天数由地震震级、台站震中距共同决定，具体对应关系如表 4.1。

表 4.1　震前正样本连续天数与地震震级、台站震中距的对应表

台站震中距 (KM)	弱震（Ms0~3）	有感地震（Ms3~4.5）	中强震（Ms4.5~6）	强震（Ms6 以上）
＜100	1 天	3 天	7 天	15 天
100~300	-	1 天	3 天	7 天
300~500	-	-	1 天	3 天

依据上述对应表，基于地震事件为各个台站生成每天一个的地震风险样本标签，具体的打标流程如图 4.1 所示。

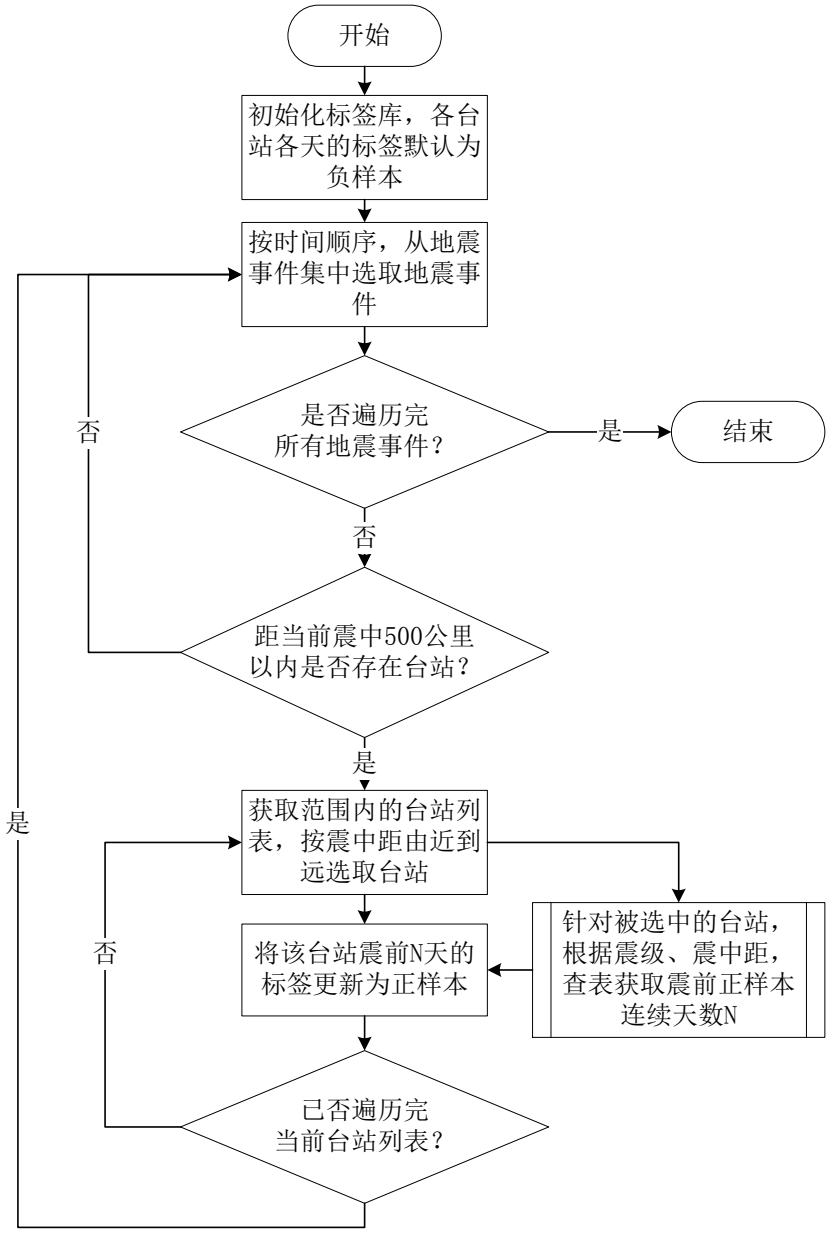


图 4.1　样本标签生成流程

4.1.2　构建特征空间

在机器学习任务中，样本通常由特征向量表示，容纳所有特征向量的向量空间称为特征空间。也就是说，特征空间是指样本所处的一个 n 维空间，n 表示特征个数。特征是个体可测量的特性或观察到的现象的特征，例如在花朵分类任务中，花朵的叶子数、花瓣的形状、颜色都可称为特征，这些特征是该分类任务的特征空间中的维度。而所有的分类算法对样本的分类，都是基于特征空间中样本的分布状况，因此构建特征空间是进行机器学习任务的重要基础。特征空间事实上是由所有特征共同决定的，而地震风险模型的特征，就是为了更好地表达近期台站附近的发震风险，而这些特征的源头，可以追溯到台站所监测到的传感器数据。

本文所使用的特征，主要分为三类：(1) 在特征提取层中基于统计方法和分形理论生成的以天为颗粒度的特征；(2) 在异常检测层中基于滑动四分位距法和 DBSCAN 密度聚类生成的以天为颗粒度的特征；(3) 项目组其他成员生成的以天为颗粒度的特征。除去第三章提及的 16 个特征外，还使用了表 4.2 所示的几个特征：

表 4.2　地震风险模型所使用的部分特征

特征名称	缩写	特征意义	理论取值范围
PCA 条带特征	PCA_B	通过滑动 PCA 提取每小时电磁信号的异常，并通过判断是否存在条带，生成条带持续天数。	0~27
电磁波形编码	EM_C	基于电磁天波形的形状，进行波形编码。	类别特征
ARIMA 差值	ARIMA_D	基于 ARIMA 预测的电磁天均值与实际值的差值。	0~12.288
电磁均值差分	EM_D	当天电磁均值与七天前电磁均值之差	0~12.288
电磁天趋势异常	EM_A	每天选择六个固定时间点的电磁数据构成一个多维数据组来描述电磁天趋势，并利用DBSCAN判断当天的电磁趋势是否出现的异常。	类别特征

4.2　分类算法的选取

在生成了地震风险模型的样本标签并构建了相应的特征空间后，需要选择合适的分类算法对近期台站附近的发震风险进行分类。在完成特征空间构建后，最终的分类效果往往取决于分类算法的选取，常用的分类算法有：基于决策树且同属集成学习的随机森林、梯度提升树；基于根据条件概率的朴素贝叶斯分类；基于高维分割的支持

向量机、神经网络等。地震风险模型的特征空间，维度不高，各特征的物理意义不同且混杂着连续特征和类别特征，而样本量也不大，因此很难发挥支持向量机、神经网络的优势。而大部分特征都源自 AETA 传感器监测信号，特征间具有一定的相关性，朴素贝叶斯分类的效果也会大大降低。而发展自决策树的随机森林和梯度提升树，适用范围广，能处理维度不高、特征间关系复杂、样本量不大的分类任务，并且算法的可解释性较强。基于 boosting 集成思想的梯度提升树拟合能力强，注重于降低偏差；基于 bagging 集成思想的随机森林具有随机选取特征、多个弱分类器集成投票等特性，注重于降低方差，更符合 AETA 数据中各台站数据特点不一、数据存在个别环境噪声的特点。综合上述原因和实验的实际效果，本文选取随机森林作为发震风险模型的分类算法。以下是对随机森林算法和上述对比算法的介绍。

4.2.1 随机森林算法

随机森林是一种集成学习算法，在训练或者预测时，会有多棵决策树基于数据集进行训练和预测，并通过投票机制集成多棵决策树的结果，作为模型的最终结果。本节首先对随机森林基分类器决策树进行介绍，其次介绍随机森林的构建。

1. 随机森林的基分类器：决策树

决策树是一种基于规则的分类算法，其分类过程逻辑清晰，实现简单，可看作一组条件判断的集合，应用于众多领域。如图 4.2 所示，典型的决策树是一种二叉树结构。树中的节点可分为根决策节点、中间决策节点和叶子结点。在决策节点上数据集会根据所选的特征属性进行分裂，而经过多次决策后完成分类的子集就是叶子结点。

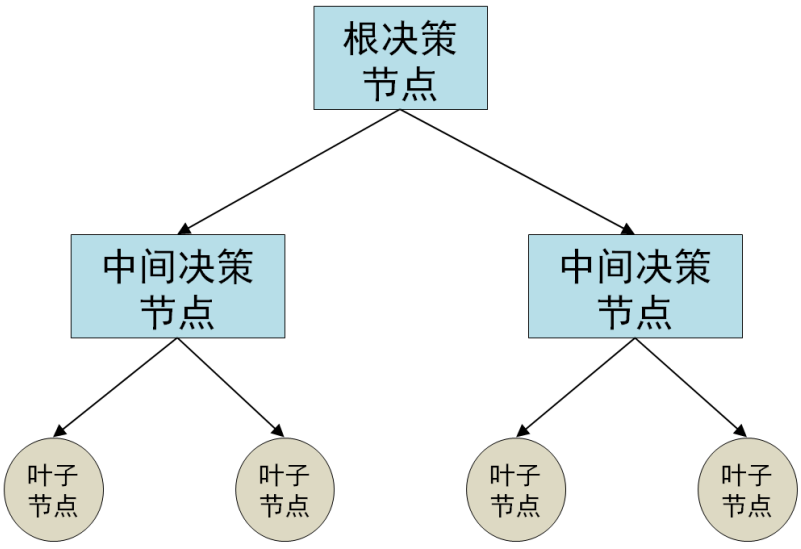


图 4.2 决策树结构

决策树的具体生成过程如下：

表 4.3　决策树生成算法

决策树生成算法
输入： TD：样本集(Train Data) F：特征集(Feature) g：最优划分特征的选择方法 输出： T：决策树 方法： 步骤 1)若 TD 非空，且 TD 中所有样本同属 C_k，则 T 为单节点树，该节点的类别标记为 C_k；若 TD 为空，则该节点的类别取父节点的多数类，返回 T； 步骤 2)若 F 为空，则 T 为单节点树，该节点的类别为 TD 中多数类 C_k，返回 T； 步骤 3)根据 g 挑选出当前的最优分裂特征 F_k； 步骤 4)基于 F_k 的每一个分裂点 g_{ki}，把 TD 划分成若干个子集 TD_i，同时初始化一个包含 i 个子节点的 T，返回 T； 步骤 5)在第 i 个子节点上，训练集取 TD_i、特征集取 F-{F_k}，递归调用步骤 1~5。

由上述决策树生成算法可知，决策树是以递归的形式对数据集进行分裂而构成的，每次分裂都是选择最优的分裂特征以及最优的分裂点，本质上是一种贪心的思想。而叶子节点停止分裂的条件有三种：

i. 样本集为空集，无法继续分裂

ii. 样本集中所有样本同属一类，继续分裂无额外增益

iii. 可用于分裂的特征集为空

最优特征的选择方法决定了决策树的生长方向和最终结构。根据特征选择方法的不同，决策树可分为基于信息增益的 ID3 决策树、基于信息增益率的 C4.5 决策树、基于 Gini 系数的 CART 决策树等。而随机森林中，所使用的基分类器是 CART 决策树。CART 的最优特征选取准则是 Gini 指标最小原则。Gini 系数的数学表达式如下：

$$Gini(S) = 1 - \sum_{i=1}^{m} {P_i}^2 \tag{4.1}$$

作为一种度量分类纯度的指标，Gini 系数和熵有相似之处，Gini 系数取值范围为 0~1，当集合内类别越杂乱，Gini 指数越接近 1。

2. 随机森林的构建

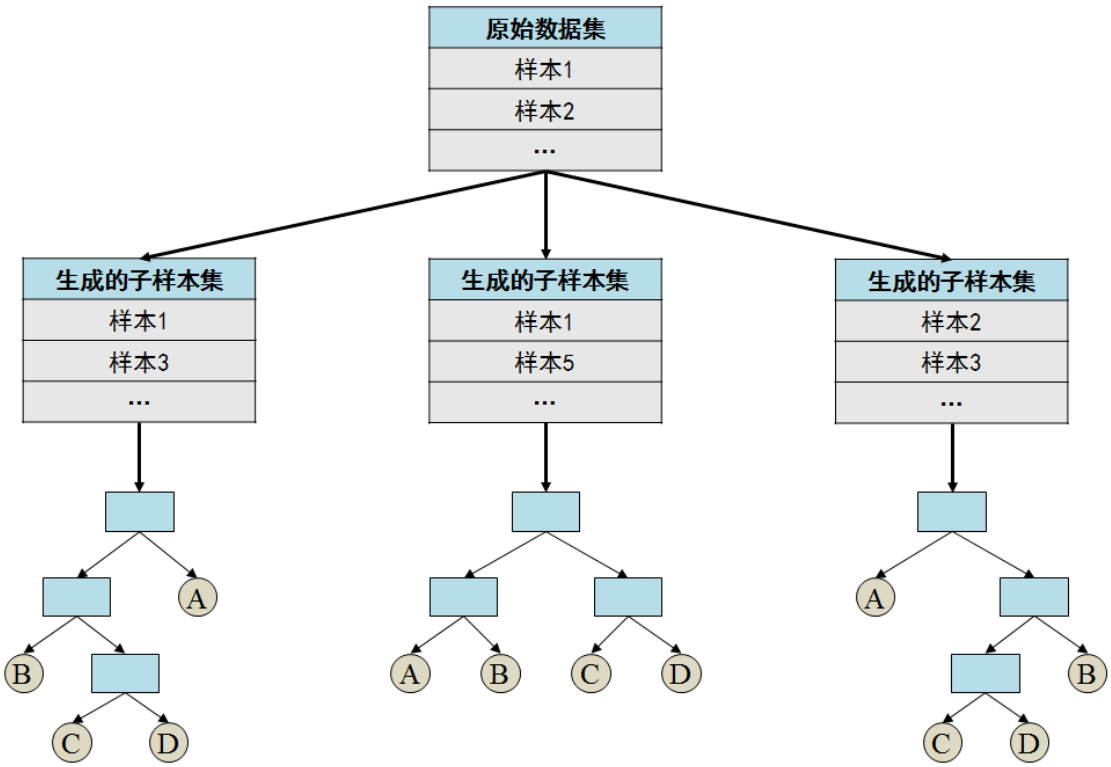


图 4.3　随机森林的构建过程

随机森林在 2001 年由 Breiman 提出，其思想是通过 bagging 和随机特征子空间，生成并集成多棵决策树，构建一个强学习器[87]。图 4.3 是随机森林的构建过程，具体包括以下三个步骤：

（1）生成子树训练集

为了保证每棵决策树的独特性，要通过随机抽样的方式从原集中为每一棵决策树分别构造各自的训练集。抽样可分为不放回和有放回两大类。采用不放回抽样时，子集的样本不会重复，据此生成的决策树多样性会更好，但不放回抽样本质上是对原集进行等分，若决策树数量较大，每棵树分得的样本就会大幅减少。有放回抽样就是在抽样后，对样本进行复制，所选样本不会从备选集中剔除，因此可以保证每棵树的样本数量，但此时同一子集中、不同子集间都可能出现相同的样本点。

在一个样本数量为 N 的集合中，每次抽样时每个样本被抽中的概率是$\frac{1}{N}$、未被抽中的概率是$(1-\frac{1}{N})$。当完成 N 次抽样时，一个样本未被抽中的概率为$(1-\frac{1}{N})^N$。$(1-\frac{1}{N})^N$可收敛于 $1/e\approx 0.368$，既当抽样次数足够多的时候，约有 36.8%的样本不会被抽中，而 1-36.8%=63.2%的样本会出现一次或多次。

（2）随机特征子空间

为每棵决策树生成了训练集之后，就到了构建决策树的过程。与普通决策树不同，随机森林中的决策树在构建过程中借鉴了随机特征子空间的思想：随机森林中的决策树在选择最优特征进行分裂时，会先从全集特征中随机选取 k 个待选特征，其次在 k 个待选特征中选择最优特征，如图 4.4 所示。通过在特征全集从随机抽取出待选特征的办法，达到随机特征子空间的效果，降低基于相似样本集生成的决策树的趋同性，进一步增加随机森林中决策树的多样性，增强随机森林的泛化能力。

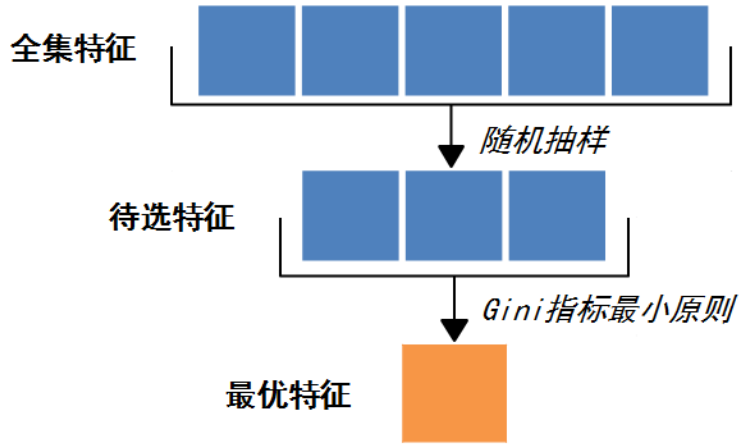


图 4.4　随机森林的子树选择分裂特征的过程

（3）基于投票机制生成随机森林

由于子样本集和子特征集的随机性，各个基分类器（即决策树）的拟合能力有所下降，但是基分类器之间的多样性得到了有力的保证。在预测时，通过投票的方式，各基分类器共同参与判决，共同判决的结果将作为随机森林算法最终的输出结果，从而实现了由多个弱学习器变为一个集成学习器（随机森林）的目标。

4.2.2　其他对比算法

1.　梯度提升树

梯度提升树归属于集成学习的 Boosting 类，是一种迭代的决策树算法，该算法与随机森林一样，是由多棵决策树组成并通过综合所有子树的结论作为最终输出。二者区别在于子树的关系，随机森林的子树属于并行关系，而梯度提升树的子树更像串行关系。梯度提升树在每一轮的迭代中，首先基于当前强学习器的结果获得本轮的损失函数，其次通过训练一个弱学习器（决策树）使本轮的残差往梯度方向减少，从而使整个集成模型向残差最少的方向学习。梯度提升树可以用于回归和分类任务，二者的算法思路一致。不同之处是，当梯度提升树用于分类时，由于样本标签是离散的，因

此无法直接获取残差，因此需要使用不同的损失函数。在二分类问题中，参考对数似然损失函数，可使用类别的预测概率值与真实概率值之差来拟合损失。

2. 朴素贝叶斯分类器

朴素贝叶斯分类器来源于统计学的两个概念：贝叶斯定理和特征条件独立假设。通过贝叶斯定理，把计算各样本分属各类的后验概率转化为对训练集先验概率和条件概率的计算；基于特征条件独立假设，假定特征之间相互独立，各特征独立影响分类结果。在上述两个基础上，朴素贝叶斯的计算方法变得高效且简单。朴素贝叶斯分类器的具体步骤如下：首先从训练集中，计算出各类别的先验概率；其次，统计各类别中各个特征的条件概率；最后计算样本分数各类的后验概率，所得概率最高的分类即分类器的预测分类。在这其中，关键是后验概率的计算，由于假设特征条件独立，则根据贝叶斯定理有如下推导：

$$\mathrm{p}(y_\mathrm{i}|x) = \frac{\mathrm{p}(x|y_\mathrm{i})\mathrm{p}(y_\mathrm{i})}{\mathrm{p}(\mathrm{x})} \tag{4.2}$$

$\mathrm{p}(\mathrm{x})$是常数项，因此只需分子取最大值$\mathrm{p}(y_\mathrm{i}|x)$就是最大值，又因为各特征条件独立，则有：

$$\begin{aligned}\mathrm{p}(x|y_\mathrm{i})\mathrm{p}(y_\mathrm{i}) &= \mathrm{p}(a_1|y_\mathrm{i})\mathrm{p}(a_2|y_\mathrm{i})\cdots\mathrm{p}(a_m|y_\mathrm{i})\,\mathrm{p}(y_\mathrm{i}) \\ &= \mathrm{p}(y_\mathrm{i})\prod_{k=1}^{n}P(a_j|y_i)\end{aligned} \tag{4.3}$$

但由于在实际应用中，贝叶斯的特征条件独立假设并不一定成立，因此模型效果并不稳定。

3. 支持向量机

支持向量机在 1995 由 Cortes 和 Vapnik 提出，是一种基于特征空间分类间隔最大化的分类器[88]。该算法的基本思想是在特征空间中寻找一个超平面，使得分类误差较小，并保证该超平面的硬间隔或软间隔最大。当样本在特征空间中不可通过超平面进行分类时，往往采用核函数等方法把原始特征投影到更高维的特征空间。由于引入核函数，支持向量机在复杂的非线性分类问题上，往往有很好的拟合能力，且得益于间隔最大原则，支持向量机的泛化能力更好且过拟合的概率更低。支持向量机被广泛应用于图像识别、自然语言处理以及基准时间序列预测检验等[89]。

4. 人工神经网络

人工神经网络借鉴了人脑神经元模型，基于信息科学的高度抽象，以节点和连接作为基本组成元素，共同组成人工神经网络。通过选择合适的神经元个数、层数和激活函数，前馈网络能无限逼近非线性函数，人工神经网络具有强大的拟合能力。随着

近年来算力和数据的爆发式增长，神经网络（特别是深度神经网络）受到了极大的关注，并在机器视觉、自然语言处理、推荐系统、工业控制等领域取得了前所未有的成功。但由于神经网络是基于传统统计中的渐进理论，对巨大样本量有天然的依赖性。而大多数情况下，样本数据有限而输入空间维度较高，此时样本数据仅是输入空间中的稀疏分布，这样的情况下人工神经网络的收敛将会异常艰难。

此外，在实际训练人工神经网络时，网络结构需要事先指定或使用启发算法在训练过程中自我修正，这都无法保证网络结构是最优的，因此网络常常陷入局部最优或直接无法收敛。同时，训练过程的优化目标是基于经验的风险最小化，但目前并没有相关的理论解析神经网络训练过程中收敛速度及方向的变化，这就导致训练效果过分依赖学习样本，不能保证网络的泛化能力。

4.3　模型的训练与效果

4.3.1　地震风险模型的训练

在地震风险模型的训练过程中，本文的主要工作包括解决数据集的样本不平衡问题和选取合适的随机森林模型超参数。

1. 针对不平衡问题，采取 SMOTE 过采样。

在实际应用中，正负样本不平衡的问题非常常见，此时预测结论往往受数据集不平衡的影响而倾向于预测为多类。对于样本不平衡，有两种解决方法：(1) 欠采样，从多类中随机抽取与少类数量接近的样本；(2) 过采样，生成更多的少类使其样本数量与多类持平。

由于本文的地震风险数据集中，正样本数量较少，若采用欠采样，会丢失负样本（无发震风险）包含的大量隐含信息，并不适合使用。因此本文采取过采样的办法来解决样本不平衡问题。最简单的过采样是对少类进行随机的有放回抽样，通过简单复制来增加少类个数，但这样做容易使模型产生过拟合。在 2002 年 Chawla 提出了 SMOTE 算法，即合成少数过采样技术，是目前学术界和工业界处理非平衡数据的常用手段[90]。SMOTE 算法的基本思想是基于原有少类样本生成模拟样本，从而增加少类样本的数量，解决类别失衡的问题。生成模拟样本的方法是 SMOTE 算法的关键，在具体实现时参考最邻近算法，首先求出各个少类样本的 K 个近邻，其次基于 K 个近邻中的任意 N 个样本进行线性插值，插值所得即模拟样本。

由于地震是一个低频事件，依据 4.1.1 小节所述方法生成的数据集中正负样本并不平衡，在 217 个 AETA 台站共计 106858 个有效样本中，正负样本比例为 1:9.3。本文采用 SMOTE 过采样技术，生成了 86100 个模拟正样本。在加入模拟正样本后，正样

本个数可达 96475，负样本个数为 96483，正负样本比例接近 1:1。

2. 选取合适的模型超参数，提升模型效果。

模型超参数是模型外部的参数，可根据具体的机器学习任务进行合理的调整。直接影响随机森林模型效果的超参数可简单分为子树超参数和森林超参数两类。重要的子树超参数包括每棵树的深度（max_depth）、分裂的标准（criterion）、叶子节点最少样本数（min_samples_leaf），这些参数共同决定了基分类器的拟合能力和对噪声的容忍程度；重要的森林超参数包括树的个数（n_estimators）、最大特征数(max_features)、是否有放回抽样（boostrap），这些参数共同决定了集成分类器的泛化能力。在调整参数过程中，本文首先根据特征的个数、样本的个数以及对地震数据的理解大致确定子树超参数，其次通过网格搜索法寻找最优的森林超参数，最后在确定森林超参数之后再微调子树超参数。各关键参数如表 4.4 所示。

表 4.4　地震风险模型中随机森林算法的超参数

超参数	特征意义	设置值
max_depth	子树的最大深度	5
criterion	树节点分裂的标准	Gini
min_samples_leaf	子树中叶子节点的最少样本数	50
n_estimators	森林中树的个数	175
max_features	子树所能选择的最大特征数	10
boostrap	是否有放回抽样	True

4.3.2　地震风险模型的分类效果

对于二分类问题，可根据预测结果和实际情况把样本分为 TP、TN、FP、FN 四类，这四类的具体含义可见表 4.5。

表 4.5　混淆矩阵

	预测正类	预测负类
实际正类	True Positives（正类判定为正类）	False Negatives（正类判定为负类）
实际负类	False Positives（负类判定为正类）	True Negatives（负类判定为负类）

为了更全面地评价地震风险模型的分类效果，本文使用准确率（accuracy）、查准率（precision）、查全率（recall）以及 AUC（Area Under Curve）这四个评价指标。

准确率是预测正确样本占总样本的比例，计算公式如下：

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.4}$$

查准率是正确预测为正类占全部预测为正类的比例，计算公式如下：

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.5}$$

查全率是正确预测为正类占全部实际正类的比例，计算公式如下：

$$\text{accuracy} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.6}$$

AUC 是指分类器 ROC 曲线（receiver operating characteristic curve）下的面积，取值范围 0~1。数值越接近 1，分类效果越好。当分类器与随机分类效果相当时，AUC=0.5。

本文从中国地震台网获取 2017 年 6 月 1 日至 2019 年 3 月 1 日间发生在中国境内及周边的 1111 次地震事件，其中弱震 361 例，有感地震 661 例，中强震 83 例，强震 6 例。依据 4.1.1 小节所述样本生成方法和 4.3.1 小节介绍的 SMOTE 过采样技术，共获得了 96475 个正样本和 96483 个负样本。使用五折交叉验证的方式对数据集共计进行了五轮的训练和验证，在每一轮中按时序抽取 146366 个样本作为训练集，36592 个样本作为验证集。

同时，为了验证随机森林的效果，本文额外训练了梯度提升树、朴素贝叶斯分类器、支持向量机、三层神经网络四个分类器进行对照实验，各分类器在五轮验证集上的平均评价指标如表 4.6 所示。

表 4.6　各分类器在五轮验证集（36592 个样本）中的平均表现

分类器	准确率	查准率	查全率	AUC
随机森林	0.531	0.540	0.423	0.696
梯度提升树	0.461	0.410	0.174	0.677
朴素贝叶斯	0.212	0.172	0.150	0.586
支持向量机	0.491	0.492	0.542	0.641
三层神经网络	0.456	0.374	0.110	0.688

五个分类器的 AUC 都大于 0.5，说明分类器的分类效果都优于随机分类。其中，使用随机森林训练出来的模型在准确率、查准率和 AUC 上都取得了最高得分。

相对于业界成熟的二分类问题，地震风险模型的 AUC 的总体水平不算高，主要原因是不同地震的前兆往往有所不同，“地震类型”众多，相对而言震例并不丰富，因此分类器在某些震例不足的“地震类型”上表现不佳，从而降低分类器的整体 AUC。

在完成对整体数据集的评估后，本文以 2018 年 10 月 31 日四川西昌 Ms5.1 地震为

例，给出基于随机森林的地震风险模型在具体震例上的预测效果。

本文以四川西昌 Ms5.1 地震附近 200 公里内所有运行台站在震前 15 天的数据作为测试集，剔除该测试集后重新训练了一个随机森林分类器，并使用该测试集进行效果检验。在 450 个测试样本（30 台站*15 天）中，分类器的准确率为 0.820、查准率为 0.743、查全率为 0.774、AUC 为 0.822。

具体的预测效果如图 4.5 所示，图中的黑色表示台站在当天出现地震风险，白色代表监测数据中没能发现地震风险，台站按震中距由近到远排列。可见，地震风险主要集中在震前七天（10 月 25 日至 10 月 31 日），且台站出现地震风险的天数随台站震中距逐渐减少越多。同时，经过数据比照，图中 10 月 17 日、10 月 21 日、10 月 24 日多个台站所出现的地震风险，应该与 10 月 17 日云南楚雄 Ms4.5 地震、10 月 21 日云南大理 Ms3.1 地震、10 月 24 日四川泸州 Ms3.2 地震有关。

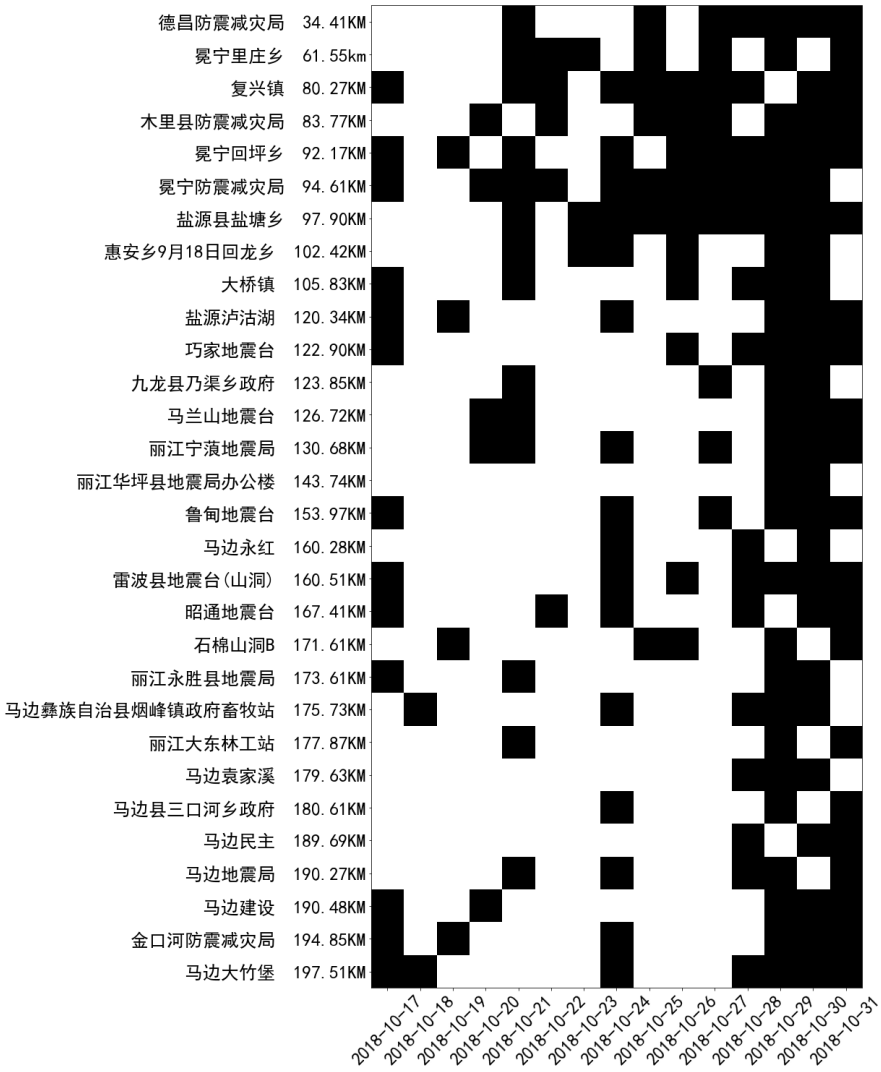


图 4.5　四川西昌 Ms5.1 地震附近台站震前 15 天的地震风险标签

4.4　本章小结

本章实现了 AETA 分层数据处理的第四层——建立地震风险模型。把经过数据预处理层、特征提取层和异常检测层处理后的台站数据输入到地震风险模型中，为台站生成一个二分类的风险标签，该标签描述在该台站附近近期有无发生地震的风险。

本章首先基于地震三要素（时间、地点、震级）生成地震风险标签，并整合特征提取层中基于统计方法和分形理论生成的特征、在异常检测层中基于滑动四分位距法和 DBSCAN 密度聚类生成的特征以及项目组其他成员生成的特征，构建了地震风险模型的特征空间。其次，通过分析算法的适用范围和地震风险模型的特征空间，从五个分类算法中选取了有较强泛化能力、拟合能力及特征兼容性的随机森林算法。再次，在模型训练过程中，使用 SMOTE 过采样解决了数据集样本不平衡的问题，并结合经验和网格搜索法为随机森林算法选取出合适的模型超参数。最后，使用五折交叉验证训练了基于随机森林算法的地震风险模型，并与由其他四种不同分类算法训练出来的模型进行对比。对比结果表明，基于随机森林算法的地震风险模型的表现最好。在总共五轮，每轮 36592 个样本的验证集中，该模型的准确率为 0.531、查准率为 0.540、查全率为 0.423、AUC 为 0.696。除此之外，本章对 AUC 相对并不高的结果进行了分析：相对丰富的“地震类型”而言，模型能学习到的震例不多，导致模型在某些震例不足的“地震类型”上表现不佳，从而降低分类器的整体 AUC。而在一些具体的地震上，模型具有更好的预测效果。以 2018 年 10 月 31 日四川西昌 Ms5.1 地震为例，在这次地震相关的 450 个样本（30 台站*15 天）中，分类器的准确率为 0.820、查准率为 0.743、查全率为 0.774、AUC 为 0.822。

第五章　预测模型的研究

在经过二到四章的分层数据处理后，AETA 数据被提炼为每个台站每天一个的地震风险标签，实现了从复杂的时序数据中提取地震风险信息的过程。从第四章的 2018 年 10 月 31 日四川西昌 Ms5.1 地震可见，地震风险在震中附近、临震前出现的概率更高，但这些地震风险标签并没有直接对地震三要素进行预测。本章基于第四章的地震风险标签，在预测地震时提出了先预测震中，再锁定预测的未来天数，最后预测震级的预测方法。并对其中的震中预测模型、未来 X 天（X 依次取 3、5、10）震级预测模型进行研究。

5.1　基于聚类算法的震中预测模型研究

在地震短临预测中，主要有两类震中预测方法：(1) 基于前兆信号的方向性和距离预测，锁定震中区域[91]。此类方法要求所监测的前兆信号具有方向性，同时信号在地下的复杂传播过程会导致距离预测的精度有限，因此震中预测的效果一般；(2) 基于电离层 TEC 异常区域，可以得出大区域内的电离层热力图，从而根据 TEC 异常直接圈定震中[92]。本文针对震中预测，提出了一个新的预测思路：基于台站的地震风险和台站经纬度，构建一个地震风险-地理位置的高维特征空间，并通过聚类的方法找出具有地震风险的簇中心，并把簇中心的地理信息提取作为震中位置。由于地震风险模型的地震风险标签是根据地震三要素设计的，因此模型在预测时所输出的地震风险标签，从某种意义上是携带距离信息的。若多个台站在基于地震风险标签构建出来的地震风险特征上表现相似时，它们在地理位置上应可聚为同簇，而簇中心对应的地理位置就是震中位置。本节首先基于各台站的地震风险标签和台站经纬度构建风险及位置特征集，其次通过聚类找出震中位置，最后给出在具体震例应用时震中预测的效果。

5.1.1　风险及位置特征集

在风险及位置特征集中，每个样本包括各台站的地震风险信息和台站经纬度两大部分。同一个特征集中各台站所使用的地震风险信息都是同一天的，在这个特征集上进行聚类可以计算出当天的震中位置。由上述可知，地震风险信息对震中位置的定位具有很大的影响，而地震风险信息是基于台站若干天的地震风险标签提取的，因此需要保证地震风险标签的质量。本章从第四章的地震风险标签集中，按照以下条件选取了合适的子集来构建风险以位置特征集：

1. 台站：一定经纬度范围内所有正常运行的台站；
2. 时间：至少连续 15 天；风险标签质量：
3. 所选取的数据集的 AUC 大于 0.80；

风险及位置特征集一共有两个位置特征和三个风险特征，具体如表 5.1 所示：

表 5.1　风险及位置特征集的特征

特征名称	缩写	特征意义	理论取值范围
台站经度	LONG	台站安装点的经度	0~180
台站纬度	LATI	台站安装点的纬度	0~90
三天风险值	RISK_3	当前台站近 3 天的地震风险天数	0~3
七天风险值	RISK_7	当前台站近 7 天的地震风险天数	0~7
十五天风险值	RISK_15	当前台站近 15 天的地震风险天数	0~15

表 5.1 中的三个风险特征中所述的地震风险天数，是统计当前台站由地震风险模型所生成的地震风险标签获得的，具体方法如图 5.1 所示。

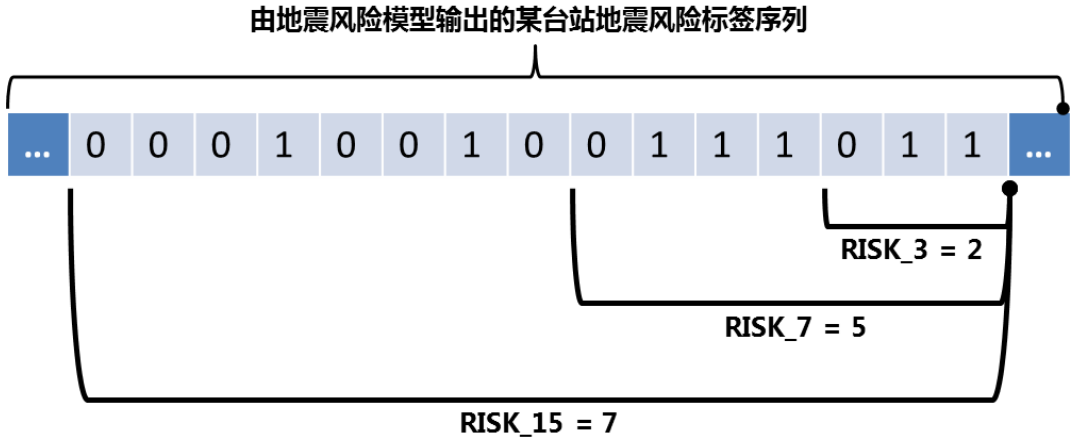


图 5.1　三个风险特征的计算方法

5.1.2　均值漂移聚类

本文假定在风险及位置特征集中，若多个台站的风险特征表现相似时，它们的地理特征应该也可聚为同簇，因此本文使用聚类算法求出这些簇中心，并把簇中心的地理信息提取作为震中位置。在选择聚类算法时，有两个要求：(1) 在这个特征集中聚类的目标是找到簇中心，因此所选聚类算法需要能计算出簇中心；(2)风险及位置特征集包含了一个区域内的多个台站，而区域内可能存在一个或多个具有发震风险的地方，即无法确认聚类的簇数，因此不能使用以 K-means 为代表的需指定聚类簇数的原型聚

类。综上，本章采用均值漂移聚类算法对风险及位置特征集中的样本进行聚类。

均值漂移算法是一种基于迭代的峰值搜索方法[93]，常用于目标跟踪、数据聚类或分类等场景。其思想类似于梯度下降方法，通过往密度梯度上升的方向不断移动，最终到达局部分布最密集的点。在介绍聚类的具体过程之前，需对漂移向量加以说明。在 d 维空间中存在 n 个样本 x_1~x_n，则基本的漂移向量为：

$$M_h(x) = \frac{1}{k_n} - \sum_{x_i \in S_h} (x_i - x) \tag{5.1}$$

其中 S_h 是一个半径为 h 的高维球体，球体内各个圆心距不同的样本 x_i 对漂移向量的贡献在式 5.1 中是一样的。为使不同距离的点对漂移向量的作用随距离增加而减少，需要引入核函数，本文采用的是常用的高斯核函数 G。调整后的漂移向量如下：

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i G(x, x_i)}{\sum_{i=1}^{n} G(x, x_i)} - x \tag{5.2}$$

表 5.2　均值漂移聚类算法

算法：均值漂移聚类算法
输入：
x_1~x_n：在 d 维空间中的待分样本
h：带宽，即高维球体的半径，本文取样本点间平均高斯距离的 1/4 作为带宽
K(x)：核函数，本文采用高斯核函数
输出：
{F}：簇中心集
{Cn}：各样本的所属类别
方法：
步骤 1) 在待分样本中随机抽取一个点，作为新的中心点 F_i；
步骤 2) 找出以 F_i 为球心，h 为半径的球体内所有的样本，记为集合 M。M 中所有样本属于 C_i 类的权重加 1，最后每个样本的分类结果取决于权重；
步骤 3) 以 F_i 为中心，依照式 5.2 计算漂移向量 m_h；
步骤 4) F_i 沿 m_h 的方向移动，移动距离是$\|m_h\|$；
步骤 5) 重复步骤 2~4，当$\|m_h\|$足够小时，可认为 F_i 已收敛。此时的 F_i 是 C_i 类的簇中心；
步骤 6)若 F_i 与某个已存在的簇中心 F_j 的距离小于阈值，则 C_i 类和 C_j 类合并；
步骤 7) 重复步骤 1~5，直到所有的点都已被标记。
步骤 8) 根据每个点在不同类上的权重，选择权重最大的类作为当前点的所属类别。

上述均值漂移聚类方法中，使用了欧式距离来表示样本点间的距离，此时各维度

对漂移向量的贡献是相同的，而在本章的风险及位置特征集中，不同特征的尺度并不相同，因此需要对各特征进行尺度变换。对于三个风险特征，考虑地震风险标签、连续累计天数以及地理位置的相关性并结合具体的实验过程，确立了变换参数。而两个位置信息中，纬度的变换是线性的，而经度的变换由于地球球体的影响，需要结合经度进行换算。各特征的具体变换方法如表 5.3 所示：

表 5.3　风险特征、位置特征的尺度变换

特征名称	缩写	变换方法
台站经度	LONG	$\mathrm{R}\times\cos^{-1}[\cos^2(\mathrm{LATI})\times\cos(\mathrm{LONG}-110)+\sin^2(\mathrm{LATI})]$
台站纬度	LATI	×111
三天风险值	RISK_3	×45
七天风险值	RISK_7	×30
十五天风险值	RISK_15	×25

5.1.3　震中预测效果

根据 5.1.1 中对地震风险标签集子集的要求，本文取纬度 31°N~33.5°N，经度 103.5°E~106.5°E 范围内的所有正常运行台站（共计 22 台）作为实验台站，获取这些台站在 2018 年 6 月 11 日至 2018 年 6 月 29 日的地震风险标签，并基于这些标签生成了 2018 年 6 月 25 日至 2018 年 6 月 29 日共计 5 天的风险及位置特征集。使用 5.1.2 的均值漂移聚类分别对这五天的风险及位置特征集进行聚类，获取每天的所有簇中心。由于台站的经纬度是经过尺度变换后再进行聚类的，因此获得的簇中心的位置特征需要经过逆变换才能得出具体的经纬度，具体如下：

$$\text{预测震中纬度}=\frac{\text{簇中心纬度}}{111} \tag{5.3}$$

$$\text{预测震中经度}=\cos^{-1}\left[\cos\left(\frac{\text{预测震中纬度}}{R}\right)-\frac{\sin^2(\text{预测震中纬度})}{\cos^2(\text{预测震中纬度})}\right]+110 \tag{5.4}$$

在风险及位置特征集中，常常会存在无发震风险的台站，而这些台站也会参与到聚类中，因此所得到的簇中心有可能是无地震风险的数据簇的中心。为了过滤这些无效震中，本文使用 5.2 小节的三个震级预测模型对震中进行投票，若有两个或以上的震级预测模型判断该震中不发震，则将该震中过滤。表 5.4 统计了 2018 年 6 月 25 日至 2018 年 6 月 29 日，实验区域内所预测的所有震中。

表 5.4　实验区域内所预测的所有震中

时间	簇中心个数	有效震中个数	有效震中的经纬度
6 月 25 日	3	-	-
6 月 26 日	3	-	-
6 月 27 日	1	1	32.52° N，105.32° E
6 月 28 日	1	1	32.77° N，104.87° E
6 月 29 日	1	1	32.63° N，104.93° E

过滤“无震”震中后得到 3 个有效震中，分别出现在 2018 年 6 月 29 日四川平武 Ms4.0 地震的震前两天、一天和发震当天。这三个预测震中与实际震中的距离分别是 80.4KM、72.4KM、61.3KM，预测震中、实际震中以及台站的位置关系如图 5.2 所示。

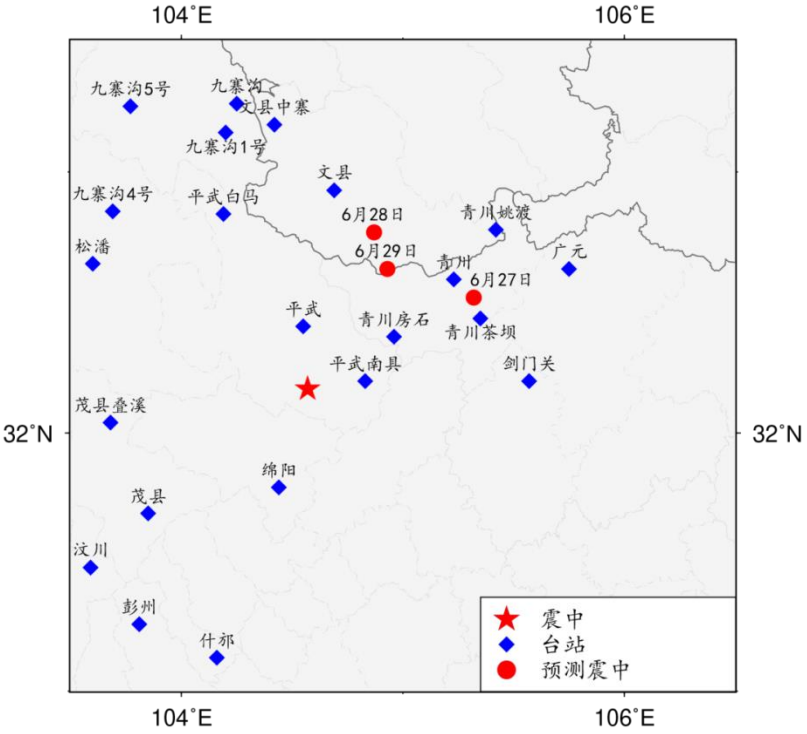


图 5.2　预测震中、实际震中以及台站的位置

周华东通过地电阻率引潮力模型研究 2007 年印尼 8.5 级地震[91]，计算所得震中距离实际震中 257KM；而基于电离层 TEC 异常所研究的众多震例，往往只能定位到一个方圆几百公里的大区域[92]。可见，本文的震中预测模型在一个高质量的风险及位置特征集中，可以取得一定的效果。但受台站布设密度所限，预测精度仍超过 50KM，若想进一步提高预测精度，需要在实验区域内布设更多台站。

5.2　基于分类算法的震级预测模型研究

震级预测在中长期预测中有不少尝试和成果，其主流方法是基于地震序列对某地区的地震震级变化进行预测[94,95]。而在地震短临预测中，震级预测主要还是依靠经验性分析，20 世纪 90 年代前后曾兴起基于模糊理论研究地震震级的热潮，主要方法是基于倾斜仪、地形变速率、氡异常等预测因子，使用模糊数学，把经验性判断转化为前兆信号与地震震级的定量或半定量关系[96–100]。但所提出的方法在往后的实际预测工作中并没有取得太大的成果。

本章结合上文的地震风险模型和震中预测模型，提出了一种新的震级预测思路：震中预测模型可以预测震中位置，实现三要素中位置的预测。而地震风险模型可以描述预测震中附近台站的地震风险。此时，震级预测模型所需要做的就是根据附近台站的地震风险推算该震中在未来数天内的可能震级。而为了进一步明确三要素中的时间，本章一共训练三个震级预测模型（下文简称为 Model_3、Model_5、Model10），分别对三天内、五天内、十天内可能出现的地震震级进行多分类预测。本章首先基于地震事件和各台站的地震风险标签建立了一个可用于多分类震级预测的数据集，其次通过多分类算法分别训练了三个震级预测模型，最后展示它们的震级预测效果。

5.2.1　基于地震事件生成震级分类数据集

震级预测模型使用的是有监督学习方法，数据集由标签和特征集两部分构成。一个震级分类样本的标签描述的是：在该位置，未来数天内发生的地震震级（类别）。由于窗口长度、分类难度都不同，三个震级预测模型的震级类别划分也略有不同，具体类别见表 5.5。

表 5.5　三个震级预测模型的预测类别

模型	震级分类方法
Model_3	Label = 0：无震；　Label = 1：(Ms0,Ms3]； Label = 2：(Ms3,Ms4.5]；　Label = 3：(Ms4.5,Ms6]； Label = 4：(Ms6, +∞]
Model_5	Label = 0：无震；　Label = 1：(Ms0,Ms4.5]； Label = 2：(Ms4.5,Ms6]；　Label = 3：(Ms6, +∞]
Model_10	Label = 0：[Ms0,Ms4.5]；　Label = 1：(Ms4.5,Ms6]； Label = 2：(Ms6, +∞]

以 Model_3 为例，说明其数据集中各类样本的选取和打标过程，具体方法如下：

1. 筛选震例构造非 0 样本。首先获取 2017 年 6 月 1 日至 2019 年 3 月 1 日期间发

生的国内地震震例集，并从中选取在震中 D 点方圆 100 公里内至少存在 1 个台站、方圆 500 公里内至少存在 3 个台站的地震事件。其次，依据地震发震时间 T 前 18 天（3 天+15 天）方圆 500 公里内各台站的地震风险标签集的质量进行筛选，筛选标准：该区域该时段内地震风险标签子集的 AUC 要大于 0.80。筛选后，符合条件的地震：强震 2 例、中强震 13 例、有感地震 41 例、弱震 40 例。

2. 构造“无震样本”。选取一个地点 D 及一个对应的时间点 T，D 和 T 满足以下要求：a) 在 T 前 3 天内，D 点方圆 150 公里内没有发生任何地震；b)在 D 点方圆 100 公里内至少存在 1 个台站、方圆 500 公里内至少存在 3 个台站；c)在时间 T 前 15 天间，D 点方圆 500 公里内所有台站构成的地震风险标签子集的 AUC 要大于 0.80。结合本章中真实震例的个数，生成 600 个(D,T)对用于构造无震事件，对应“震级”为 0。

3. 从 1 和 2 中，得到了 96 个真实震例和 600 个生成的无震样本，而这 96 个真实震例还需要扩充为 288 个样本（96*3）。以 2018 年 11 月 26 日台湾海峡 Ms6.2 地震为例，地点 D 为 (23.28°N，118.6°E)，发震时间 T 前三天（T-3，T-2，T-1），可生成三个 Label=4（震级大于 Ms6.0）的样本。这三个样本描述的是，在地点 D，未来三天会有一个震级大于 6 级的地震。

4. 用于该震级预测模型的震级分类样本完成，一共有 5 类样本，样本总数为 888。

Model_5、Model_10 的样本打标方法和构建过程类似，只需修改如下两点：1.把发震时间 T 前三天改为震前五天、震前十天。所对应生成的真实震例样本数分别为 480、960，对应的无震样本分别为 1000、2000；2.把样本类别，从 5 类修改为 4 类和 3 类，依照各自的震级分类标准进行打标即可。

完成打标后的震级分类样本都包含具体时间 T、具体地点 D 的对应信息。基于对应时间 T 前后、地点 D 附近的台站的地震风险标签集，本章构建两大类震级分类特征：范围内台站地震风险值的平均值、范围内地震风险值超过设置阈值的台站占比。其中，各台站的地震风险值使用 5.1.1 风险及位置特征集小节中的 RISK_3、RISK_7、RISK_15。对台站到地点 D 的距离范围、地震风险值进行组合后，共有 3*3*2=18 个特征，具体的特征以及生成特征所需的阈值如表 5.6 所示。

表 5.6　震级预测模型所用的震级分类特征

距离范围	台站地震风险值	特征缩写	特征含义
0~100KM	RISK_3	AVG_100_3	100KM 内所有台站 RISK_3 的均值
		PROP_100_3	100KM 内 RISK_3≥2 的台站占比
	RISK_7	AVG_100_7	100KM 内所有台站 RISK_7 的均值
		PROP_100_7	100KM 内 RISK_7≥4 的台站占比
	RISK_15	AVG_100_15	100KM 内所有台站 RISK_15 的均值
		PROP_100_15	100KM 内 RISK_15≥8 的台站占比
100~300KM	RISK_3	AVG_300_3	100KM~300KM 内所有台站 RISK_3 的均值
		PROP_300_3	100KM~300KM 内 RISK_3≥1 的台站占比
	RISK_7	AVG_300_7	100KM~300KM 内所有台站 RISK_7 的均值
		PROP_300_7	100KM~300KM 内 RISK_7≥2 的台站占比
	RISK_15	AVG_300_15	100KM~300KM 内所有台站 RISK_15 的均值
		PROP_300_15	100KM~300KM 内 RISK_15≥4 的台站占比
300~500KM	RISK_3	AVG_500_3	300KM~500KM 内所有台站 RISK_3 的均值
		PROP_500_3	300KM~500KM 内 RISK_3≥0 的台站占比
	RISK_7	AVG_500_7	300KM~500KM 内所有台站 RISK_7 的均值
		PROP_500_7	300KM~500KM 内 RISK_7≥1 的台站占比
	RISK_15	AVG_500_15	300KM~500KM 内所有台站 RISK_15 的均值
		PROP_500_15	300KM~500KM 内 RISK_15≥2 的台站占比

5.2.2　模型训练与震级预测效果

本节首先使用主成分分析法对震级分类数据集进行降维；其次使用不同的算法训练震级预测模型 Model_3，通过对比效果选择支持向量机作为震级预测模型的分类算法；最后，利用验证集和具体震例分析 Model_3、Model_5、Model_10 的预测效果。

表 5.6 中的 18 个特征都源自 D 点 500 公内台站的地震风险值，这些特征的相关性较大，特征集存在大量冗余信息。同时在高维空间中样本的采样密度低，不利于进行模型训练。因此，需采用主成分分析对特征集进行降维，降维后样本在特征空间的采样率密度增大，同时也带来了降噪的额外收益。PCA 需要指定重构维数，经实验，本节的重构维数 k=6，此时在数据集中重构误差小于 0.01，超过 99%的信息被保留。

经过 PCA 降维后，数据集只有 6 个特征，随机森林、梯度提升树等集成学习方法并不适用。本节采用决策树、朴素贝叶斯、支持向量机等三种算法，使用五折交叉验证的方式对震级预测模型进行训练。以 Model_3 为例，使用不同算法训练的模型在五个验证集中的平均表现如表 5.7 所示。

表 5.7　不同算法生成的 Model_3 在验证集中的平均表现

分类器	宏平均查准率	宏平均查全率	宏平均 F 值
决策树	0.799	0.766	0.778
朴素贝叶斯	0.840	0.783	0.803
支持向量机	0.873	0.852	0.860

从表 5.7 可以看到，使用支持向量机训练的模型表现最好。因此，三个震级预测模型均选用支持向量机进行训练。为了更完整地分析模型的效果，给出 Model_3、Model_5 和 Model_10 在各个类别的具体分类效果以及混淆矩阵，具体可见表 5.8~表 5.11。

表 5.8　三个震级预测模型的具体分类效果

模型	震级类别	查准率	查全率	F1 值	样本数
Model_3	无震	0.955	0.925	0.94	600
	(Ms0,Ms3]	0.708	0.708	0.708	120
	(Ms3,Ms4.5]	0.745	0.667	0.704	123
	(Ms4.5,Ms6]	0.625	0.769	0.69	39
	(Ms6,+∞]	0.172	0.833	0.286	6
	宏平均	**0.873**	**0.852**	**0.86**	**888**
Model_5	无震	0.93	0.92	0.925	1000
	(Ms0,Ms4.5]	0.866	0.751	0.804	405
	(Ms4.5,Ms6]	0.462	0.662	0.544	65
	(Ms6,+∞]	0.128	0.6	0.211	10
	宏平均	**0.887**	**0.86**	**0.87**	**1480**
Model_10	(Ms0,Ms4.5]	0.979	0.967	0.973	2810
	(Ms4.5,Ms6]	0.505	0.408	0.451	130
	(Ms6,+∞]	0.11	0.45	0.176	20
	宏平均	**0.953**	**0.939**	**0.945**	**2960**

表 5.9　Model_3 的混淆矩阵

	无震	**(Ms0,Ms3]**	**(Ms3,Ms4.5]**	**(Ms4.5,Ms6]**	**(Ms6,+∞]**
无震	**490**	33	28	27	22
(Ms0,Ms3]	16	**95**	5	3	1
(Ms3,Ms4.5]	4	28	**82**	6	3
(Ms4.5,Ms6]	0	1	10	**24**	4
(Ms6,+∞]	0	0	0	2	**4**

表 5.10　Model_5 的混淆矩阵

	无震	**(Ms0,Ms4.5]**	**(Ms4.5,Ms6]**	**(Ms6,+∞]**
无震	**904**	34	32	30
(Ms0,Ms4.5]	113	**224**	36	32
(Ms4.5,Ms6]	2	30	**27**	6
(Ms6,+∞]	0	0	6	**4**

表 5.11　Model_10 的混淆矩阵

	[Ms0,Ms4.5]	**(Ms4.5,Ms6]**	**(Ms6,+∞]**
[Ms0,Ms4.5]	**2683**	68	59
(Ms4.5,Ms6]	89	**0**	41
(Ms6,+∞]	2	15	**3**

注：表 5.9、表 5.10、表 5.11 的横轴是预测值，纵轴是实际值。

如表 5.8 所示，提前越多天数进行预测的模型，其分类结果的宏平均指标越好。这主要得益于模型分类的类别数，类别越少分类难度也小。而在相同类别中 Model_3 的分类效果最好，如在(Ms4.5,Ms6]这类上，Model_3 的查准率和查全率分别是 0.625 和 0.769。此外，综合震级划分的精细度和分类效果，Model_3 的效果最好，但由于其提前量在三个模型中最小，因此在实在预测时可以综合三个模型的预测效果。

中国地震台网中心的姚丽统计了基于地磁低点位移法进行地震短临预测的效果，在 2008 年 1 月至 2016 年 5 月期间，在中国大陆区域共发生了 246 次 5 级以上地震，准确率为 0.16[101] 。

在完成对整体数据集（96 个真实震例所生成的有震样本和按比例生成的无震样本）的评估后，本文以 2018 年 10 月 31 日四川西昌 Ms5.1 地震作为具体震例，展示三个震

级预测分类器在此地震前 10 天的预测效果。为避免数据泄露，从训练集中剔除该震例所生成的样本，在此基础上重新训练了三个震级预测分类器，并使用该震例的相关样本进行效果检验。在这个数据集中，Model_3、 Model_5、Model_10 的准确率分别为 0.5，0.6，0.2。详细的预测结果如图 5.3 所示，图中蓝色为 Model_3 的预测值，红色为 Model_5 的预测值，黄色是 Model_10 的预测值。在震前 8、9 天的时候，Model_3 和 Model_5 都出现了地震预报，结合两个模型，预测震级应为 Ms0~Ms3，经查实该地震属于误报。而在震前两天，Model_3 预报震级为 Ms3~Ms4.5、Model5 预报震级为 Ms4.5~Ms6、Model_10 预报为 Ms4.5~Ms6；在震前一天，三个模型的预报震级都是 Ms4.5~Ms6。该震例证明，尽管三个模型在震前十天的准确率不高，但结合三个模型的预测结果，在发震前两天可以大致锁定该地震的震级，从而在一定程度上起到了震级预测的作用。

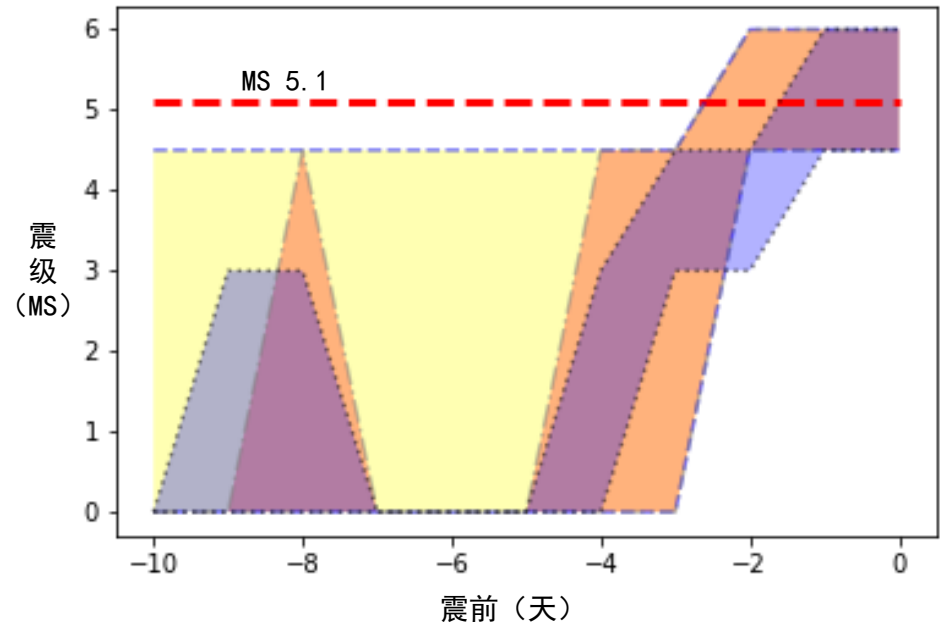


图 5.3　Model_3、 Model_5、Model_10 在西昌 Ms5.1 地震前 10 天的震级预测效果

5.3　本章小结

本章在第四章生成的地震风险标签集的基础上，提出了对地震三要素按震中、发震时间、震级顺序进行独立预测的方法，并分别对震中预测模型和震级预测模型进行了研究。

本章首先对震中预测模型进行了研究。震中预测模型是基于以下猜想提出的：若多个台站有相似的地震风险特征，则它们在地理位置上应可聚为同簇。5.1.1 小节中，

基于各台站的地震风险标签和台站经纬度，构建了风险及位置特征集；其次在 5.1.2 小节，研究了均值漂移聚类算法并针对风险及位置特征集设计了具体的特征变换方法；最后，在纬度 31°N~33.5°N，经度 103.5°E~106.5°E 范围内，研究 2018 年 6 月 25 日至 2018 年 6 月 29 日的震中预测情况，通过均值漂移聚类从风险及位置特征集中得到多个预测震中，而震前两天至发震当天所预测的震中与 2018 年 6 月 29 日四川平武 Ms4.0 地震的震中分别相距 80.4KM、72.4KM、61.3KM。

其次，本章对震级预测模型进行了研究。震级预测模型的研究有两个重要前提：(1)第四章的地震风险标签集能较好地描述台站的地震风险；(2) 震中预测模型能较好地实现对震中的预测。在这个假设前提下，震中预测模型专注于对固定位置未来数天可能出现的地震进行震级预报。根据预测提前天数的不同，本章共训练三个震级预测模型，分别对未来三天内、未来五天内、未来十天内可能出现的地震震级进行多分类预测，即 Model_3、Model_5 和 Model_10。5.2.1 小节详细说明了如何根据地震事件和地震风险数据集，构建不同分类的样本。在 5.2.2 小节中，首先使用主成分分析 PCA 对震级分类数据集进行了降维；其次使用不同的算法训练 Model_3，通过对比效果，选择了支持向量机作为震级预测模型的分类算法；再次，展示了 Model_3、Model_5、Model_10 在验证集上的具体分类效果，并综合类别精细度和分类效果对三个模型的表现进行了分析，发现在宏平均指标表现最差的 Model_3 在具体分类中效果最好；最后，以 2018 年 10 月 31 日四川西昌 Ms5.1 地震为例，展示了三个模型在震前十天的预测效果，发现在震前两天可以大致锁定该地震的震级为 Ms4.5~Ms6，能起到一定程度的震级预测作用。

基于本章的研究成果，在进行地震预测实践时可通过以下步骤进行地震预测：(1)使用震中预测模型，在由地震风险标签及台站信息表构成的风险及位置特征集上通过均值漂移聚类预测出震中的具体位置；(2)根据预测的震中位置及其周围台站的风险特征，分别使用三个震级预测模型预测未来 3 天、5 天、10 天可能出现的地震震级；(3)综合震中预测模型和三个震级预测模型的结果，得出预测的震中以及未来 3 天、5 天、10 天可能出现的地震震级。

本章所研究的震中预测模型和震级预测模型都取得了一定的效果，但这是基于各台站的地震风险标签质量足够高的情况下所做的研究。由于本章所研究的震中预测模型、震级预测模型和第四章的地震风险模型是一个流式处理的关系，若地震风险模型效果不佳，误差的传导会使震中预测模型难以锁定震中，震级预测模型亦将无法找不到具体的分析区域从而无法进行震级预测。但本章所提的两个模型，确实为研究震中预测、震级预测提供了全新的思路，具有一定的研究价值。

第六章　总结与展望

6.1　总结

地震作为一种具有强大破坏力的自然灾害，每年所造成的死亡人数占自然灾害导致的死亡人数总数的 54%，其带来的巨大伤害也激起了人类对实现地震预测的热切追求。如今，国内外地震预测研究与地震预报实践的水平仍然较低，特别是更具实用价值的短临预测还远远无法满足社会需求。为了探索地震预测这一世界难题，北京大学深圳研究生院集成微系统重点实验室研制了多分量地震监测系统 AETA，给地震预测工作带来了数据基础。本文所做的主要工作如下：

1. 提出并实现了 AETA 分层数据处理。在数据预处理层，完成了对断电数据的自动化识别和清洗、对缺失值进行分类填补、重新统一了 AETA 数据的传输采样率；在特征提取层，基于统计方法提取了多种统计特征、基于分形理论提取了 AETA 电磁扰动数据的分形维数；在异常检测层，基于滑动四分位距法计算一维数据的异常值、基于 DBSCAN 密度聚类对多维数据进行异常点提取；在地震风险模型层，基于地震三要素生成了地震风险标签、结合特征序列与异常序列构建了特征空间、基于随机森林算法训练了地震风险模型。模型在五折交叉检验的验证集（36592 个样本）上的准确率为 0.531、查准率为 0.540、查全率为 0.423、AUC 为 0.696。通过分析，本文认为 AUC 不高的原因是：震例个数相对于丰富的“地震类型”而言并不充足，导致模型在某些震例较少的“地震类型”上表现不佳，从而降低分类器的整体 AUC。在此之外，为了研究地震风险模型在具体震例上的效果，本文以 2018 年 10 月 31 日四川西昌 Ms5.1 地震为例，发现在这次地震相关的 450 个样本中，分类器的准确率为 0.820、查准率为 0.743、查全率为 0.774、AUC 为 0.822。

2. 提出并实现了对地震三要素进行独立预测的方法。在调研了现有的地震预测方法后，确立了对地震三要素按照地点、时间、震级的顺序进行独立预测的基本研究思路。在 AETA 分层数据处理所生成的地震风险标签集上，提出并实现了一个震中预测模型和三个震级预测模型。在震中预测模型中，基于台站的地震风险标签集和台站经纬度，构建了风险及位置特征集，根据经纬度的距离换算和地震风险特征的尺度换算对特征集进行了特征变换，通过均值漂移聚类找出具有地震风险的簇中心，并从簇中心提取出震中的具体位置。在纬度 31°N~33.5°N，经度 103.5°E~106.5°E 范围内，研究了 2018 年 6 月 25 日至 2018 年 6 月 29 日的震中预测情况，震前两天至发震当天共计算出三个预测震中，这三个预测震中与 2018 年 6 月 29 日四川平武 Ms4.0 地震的

震中分别相距 80.4KM、72.4KM、61.3KM。本文所建立的三个震级预测模型（Model_3、Model_5、Model10）分别对固定位置未来三、五、十天可能出现的地震震级进行多分类预测。在构建时，首先基于地震事件和各台站的地震风险特征构建了用于多分类震级预测的数据集，其次使用主成分分析 PCA 对数据集进行了降维并选用了效果最好的支持向量机训练了 Model_3、Model_5、Model_10。在验证集上，三个模型的宏平均查准率可达 0.873、0.887、0.953，宏平均查全率可达 0.852、0.860、0.939，但这主要归功于对无震类别的优秀判断。在对真实地震样本进行判断时，Model_3 在 Ms3~Ms4.5 的查准率、查全率可达 0.745、0.667，而其余模型的表现都低于 0.5。尽管查准率和查全率不高，但在分析具体震例时发现，综合三个模型后所得出的震级预测具有一定的准确性。以 2018 年 10 月 31 日四川西昌 Ms5.1 地震为例，展示了三个模型在震前十天的预测效果，发现在震前两天可以大致锁定该地震的震级为 Ms4.5~Ms6，能起到一定程度的震级预测作用。

本文的工作，为预测实践提供了一个新的方法：首先利用 AETA 分层数据处理计算各台站的地震风险标签，其次通过震中预测模型求得可能震中，最后结合三个震级预测模型判断未来数天内震中位置可能出现的地震的震级。本文所完成的 AETA 分层数据处理框架可以实现从 AETA 传感器时序数距到地震风险标签的转化，而所完成的震中预测、震级预测在基于质量较高(AUC>0.8)的地震风险标签集中，取得了良好的地震预测效果。

6.2　展望

本文完成了 AETA 分层数据处理、震中预测模型和震级预测模型，实现了从 AETA 数据出发，对地震三要素进行预测。但由于所研究的震中预测模型、震级预测模型和 AETA 分层数据处理是一个流式处理的关系，当 AETA 分层数据处理所获得的台站地震风险标签效果不佳时，误差的传导会使震中预测模型的预测精度下降，而在虚假的震中进行震级预测意义并不大。因此，未来的主要工作是提高 AETA 的分层数据处理的效果，并在此基础上对震中预测模型和震级预测模型进行调优。为实现这个目标，需要：

1. 保持 AETA 系统的长期稳定运行，在监测范围内获取更多地震震例。同时，考虑升级 AETA，增加监测的信号，丰富数据来源。只有通过增加样本、提升特征的信息含量，才能从根本上提高特征空间的样本采样率，才能让地震风险模型覆盖到更多类型的地震，提高地震风险模型的泛化能力。

2. 基于 AETA 分层数据处理框架的可扩展性来提升框架识别地震风险的能力。如，

在数据预处理层，可通过获取如天气、磁暴等外部信息源，对 AETA 数据进行降噪；在特征提取层，进一步挖掘 AETA 原始数据；在异常检测层，可以针对时间序列的特点实现更多的异常检测方法，例如基于时序数据的特征去提取 SRSS 波的异常变化；在特征提取层和异常检测层获取了足够多的特征后，可以尝试把地震风险模型从二分类模型变为拟合风险指数的回归模型。

3. 在震中预测方面，要想从根本上提升定位精度，需要加密 AETA 台站的布设；而在算法层面，可以进一步开展对风险-位置这个高维特征空间的研究，构建更加贴合地震活动的风险特征。

4. 在震级预测方面，效果主要依赖于震中的准确和附近台站风险指标的效果。针对震级预测模型本身，可以在震例更多、监测数据更丰富的情况下，把震级进行更精细的划分，提高震级分辨率。

参考文献

[1] 王根龙, 张军慧. 中国地震灾害防御对策的进展及今后的发展趋势[J]. 防灾科技学院学报, 2004, 6(2): 17－19.

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

[3] 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.

[4] 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.

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

[6] Nanjo K Z, Yoshida A. A b map implying the first eastern rupture of the Nankai Trough earthquakes[J]. Nature Communications, 2018, 9(1): 1117.

[7] 路鹏, 李志雄, 陶本藻等. 震级频度与古登堡-里克特关系式偏离的前兆意义[J]. 地震, 2006, 26(4): 1－8.

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

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

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

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

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

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

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

[15] 房立华, 吴建平, 苏金蓉等. 四川九寨沟 M_s7.0 地震主震及其余震序列精定位[J]. 科学通报, 2018(7).

[16] 尹鹏, 张永志, 焦佳爽等. 基于 GRACE 数据的尼泊尔 Ms8.1 地震北向重力梯度变化[J]. 地震学报, 2018, 40(1).

[17] Qiu Q, Moore J D P, Barbot S, et al. Transient rheology of the Sumatran mantle wedge revealed by a decade of great earthquakes[J]. Nature Communications, 2018, 9(1): 995.

[18]尹祥础, 尹灿. 非线性系统失稳的前兆与地震预报——响应比理论及其应用[J]. 中国科学化学：中国科学, 1991, 21(5): 512－518.

[19] Chen L, Kong X, Zheng Z, et al. Application of geometric moving average martingale algorithm in anomaly analysis before earthquake based on sliding window[J]. Journal of Computer Applications, 2013, 33(12): 3608–3610.

[20] Hasbi A M, Momani M A, Ali M A M, et al. Ionospheric and geomagnetic disturbances during the 2005 Sumatran earthquakes[J]. Journal of Atmospheric and Solar-Terrestrial Physics, 2009, 71(17): 1992–2005.

[21] Strigunova M S, Shurygin A M. Sliding-Window Scalar Multiplication of Matrices and Earthquake Prediction[J]. Automation & Remote Control, 2004, 65(7): 1059–1065.

[22] Holtzman B K, Paté A, Paisley J, et al. Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field[J]. Science Advances, 2018, 4(5): eaao2929.

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

[24] Bernardi A, Fraser A C. Low-frequency magnetic field measurements near the epicenter of the M s 7.1 Loma Prieta Earthquake[M]. 1990.

[25] Hirano T, Hattori K. ULF geomagnetic changes possibly associated with the 2008 Iwate–Miyagi Nairiku earthquake[J]. Journal of Asian Earth Sciences, 2011, 41(4): 442–449.

[26] Gokhberg M B, Morgounov V A, Yoshino T, et al. Experimental measurement of electromagnetic emissions possibly related to earthquakes in Japan[J]. Journal of Geophysical Research Solid Earth, 1982, 87(B9): 7824–7828.

[27] 赵国泽, 汤吉, 邓前辉等. 人工源超低频电磁波技术及在首都圈地区的测量研究[J]. 地学前缘, 2003, 10(s1): 248－257.

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

[29] 郭自强, 尤峻汉, 李高等. 破裂岩石的电子发射与压缩原子模型[J]. 地球物理学报, 1989, 32(2): 173－177.

[30] 包德修, 和仁道, 马伟林等. 地震电磁信息的偶电体模型[J]. 中国地震, 1991(4): 83－86.

[31] 钱书清, 吕智. 地震电磁辐射前兆不同步现象物理机制的实验研究[J]. 地震学报, 1998(5): 535－540.

[32] 袁运斌, 欧吉坤. 利用 IGS 的 GPS 资料确定全球电离层 TEC 的初步结果与分析[J]. 自然科学进展, 2003, 13(8): 885－888.

[33] 万卫星, 宁百齐, 刘立波等. 中国电离层 TEC 现报系统[J]. 地球物理学进展, 2007, 22(4): 1040－1045.

[34] 余涛, 万卫星, 刘立波等. 利用 IGS 数据分析全球 TEC 的周年和半年变化特性[J]. 地球物理学报, 2006, 49(4): 943－949.

[35] 张东和, 萧佐. 利用 GPS 计算 TEC 的方法及其对电离层扰动的观测[J]. 地球物理学报, 2000, 43(4): 451－458.

[36] 郭子祺, 郭自强. 岩石破裂中多裂纹辐射模型[J]. 地球物理学报, 1999, 42(s1): 172－177.

[37] 陈衍景, 李超, 张静等. 秦岭钼矿带斑岩体锶氧同位素特征与岩石成因机制和类型[J]. 中国科学:地球科学, 2000, 30(s1): 64－72.

[38] 阮百尧, 徐世浙. 三维地面断面电阻率测深有限元数值模拟[J]. 地球科学, 2001, 26(1): 73－77.

[39] 郭自强, 郭子祺, 钱书清等. 岩石破裂中的电声效应[J]. 地球物理学报, 1999, 42(1): 74－83.

[40] 刘君, 安张辉, 范莹莹等. 芦山 M_S7.0 与岷县漳县 M_S6.6 地震前电磁扰动异常变化[J]. 地震, 2015, 35(4): 43－52.

[41] 解滔, 刘杰, 卢军等. 2008 年汶川 M_S8.0 地震前定点观测电磁异常回溯性分析[J]. 地球物理学报, 2018, 61(05): 1922－1937.

[42] 高曙德, 汤吉, 杜学彬等. 汶川 8.0 级地震前后电磁场的变化特征[J]. 地球物理学报, 2010, 53(3): 512－525.

[43] 丁跃军. 电磁扰动原理与地震[C]//中国科学技术协会 2008 防灾减灾论坛论文集. 2008.

[44] 丁鉴海, 黄雪香, 戴淑玲. 地震活动的月相效应[J]. 地震, 1994(4): 7－13.

[45] 吕桂芳. 1988 年澜沧—耿马地震前地磁异常变化特征[J]. 地震地磁观测与研究, 1997(2): 53－59.

[46] 卢振业, 孙若昧, 邢如英. 唐山地震前后地磁 Z 分量功率谱异常[J]. 地震, 1983(2): 6－10.

[47] Shechtman E, Irani M. Space-time behavior based correlation[C]//IEEE Computer Society Conference on Computer Vision & Pattern Recognition. 2005.

[48] 冯志生, 李鸿宇, 张秀霞等. 地磁谐波振幅比异常与强地震[J]. 华南地震, 2013, 33(3): 9－15.

[49] 孙若昧, 卢振业. 唐山地震前后地磁 Z 分量功率谱异常性质的探讨[J]. 中国地震, 1985(4):50-54.

[50] Amezquita-Sanchez J P, Adeli H. Signal Processing Techniques for Vibration-Based Health Monitoring of Smart Structures[J]. Archives of Computational Methods in Engineering, 2016, 23(1): 1–15.

[51] Alperovich L, Zheludev V. Wavelet transform as a tool for detection of geomagnetic precursors of earthquakes[J]. Physics & Chemistry of the Earth, 1998, 23(9–10): 965–967.

[52] 林云芳, 曾小苹, 续春荣等. 地磁方法在地震预报中的应用[J]. 地震地磁观测与研究, 1999, 20(6): 35－44.

[53] 陈伯舫. 日本鹿屋台地磁转换函数的变化[J]. 华南地震, 2003, 23(1): 8－12.

[54] 陈化然, 杜爱民, 王亚丽等. 地磁低点位移与地磁场等效电流体系关系的初步研究[J]. 地震学报, 2009, 31(1): 59－67.

[55] 丁鉴海, 余素荣, 肖武军. 地磁"低点位移"现象与昆仑山口西 8.1 级地震[J]. 地震工程学报, 2003, 25(1): 16－21.

[56] Hattori K, Serita A, Yoshino C, et al. Singular spectral analysis and principal component analysis for signal discrimination of ULF geomagnetic data associated with 2000 Izu Island Earthquake Swarm[J]. Physics & Chemistry of the Earth Parts A/b/c, 2006, 31(4–9): 281–291.

[57] Telesca L, Lapenna V, Vallianatos F, et al. Multifractal features in short-term time dynamics of ULF geomagnetic field measured in Crete, Greece[J]. Chaos Solitons & Fractals, 2004, 21(2): 273–282.

[58] Hayakawa M, Ito T, Smirnova N. Fractal analysis of ULF geomagnetic data associated with the Guam Earthquake on August 8, 1993[J]. Geophys.res.lett, 1999, 26(18): 2797–2800.

[59] Gotoh K, Hayakawa M, Smirnova N A, et al. Fractal analysis of seismogenic ULF emissions[J]. Physics & Chemistry of the Earth, 2004, 29(4): 419–424.

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

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

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

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

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

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

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

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

[68] 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.

[69] 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.

[70] 陈斌, 陈松灿, 潘志松等. 异常检测综述[J]. 山东大学学报(工学版), 2009, 39(06): 13－23.

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

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

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

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

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

[76] 蔡爱民, 查良松. 基于分形理论的安徽省旱、洪涝灾害时序特征分析[J]. 安徽农业大学学报, 2005, 32(4): 546－550.

[77]孙金花, 冯英浚, 胡健. 基于分形理论的股票时序数据离群模式挖掘研究[J]. 运筹与管理, 2008, 17(5): 135－140.

[78] 熊正丰. 金融时间序列分形维估计的小波方法[J]. 系统工程理论与实践, 2002, 22(12): 48－53.

[79] Esteller R, Vachtsevanos G, Echauz J, et al. A Comparison of waveform fractal dimension algorithms[J]. IEEE Trans Circuits Syst, 2001, 48(2): 177–183.

[80] Grassberger P, Procaccia I. Characterization of Strange Attractors[J]. Physical Review Letters, 1983, 50(5): 346.

[81] Procaccia I. Measuring the strangeness of strange attractors[J]. Physica D Nonlinear Phenomena, 1983, 9(1): 189–208.

[82]秦建强, 孔祥玉, 胡绍林等. 一维时间序列分形维数算法对比分析[J]. 计算机工程与应用, 2016, 52(22): 33－38.

[83] 吕玉红. 时间序列异常检测算法的研究与应用[D]. 电子科技大学, 2018.

[84] Brys G, Hubert M, Struyf A. A Robust Measure of Skewness[J]. Journal of Computational & Graphical Statistics, 2004, 13(4): 996–1017.

[85] 张明敏, 刘智敏, 刘盼等. 九寨沟7.0级地震前电离层TEC异常分析[J]. 测绘工程, 2018, 27(12): 24－30.

[86] Rich F J, Sultan P J, Burke W J. The 27-Day Variations of Plasma Densities and Temperatures in the Topside Ionosphere[J]. Journal of Geophysical Research Space Physics, 2003, 108(A7).

[87] Cutler A, Cutler D R, Stevens J R. Random Forests[J]. Machine Learning, 2004, 45(1): 157–176.

[88] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273–297.

[89] 韩家炜. 数据挖掘概念与技术[M]. 2005.

[90] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.

[91] 周华东. 地电阻率地震预报简述与地震震中距的确定[D]. 中国科学技术大学, 2010.

[92] 张小涛, 宋治平, 李纲. 九寨沟 M_S7.0 地震的前兆异常时空演化特征及其分析[J]. 中国地震, 2018, 34(04): 772－780.

[93] Comaniciu D, Meer P. Mean shift analysis and applications[C]//Iccv. 1999.

[94] 张桂铭, 刘文锋. 基于震例研究的地震预测预报分析[J]. 中国地震, 2013, 29(4): 528－536.

[95] 蔡静观, 刘正荣. 未来地震震级的定量计算[J]. 地震研究, 1989(3): 219－225.

[96] 郑熙铭, 冯德益. 地形变速率与地震震级的模糊判别[J]. 华北地震科学, 1987(S1): 306－312.

[97] 谷懿, 葛良全, 王广西等. 汶川地震震后大成都地区断裂带活动性氡气测量分析评价[J]. 工程地质学报, 2009, 17(3): 296－300.

[98] 杜建国, 宇文欣, 李圣强等. 八宝山断裂带逸出氡的地球化学特征及其映震效能[J]. 地震, 1998(2): 155－162.

[99] 陈立军, 许峻, 陈晓逢等. 用块体应变能预测强震位置与震级的研究[J]. 地震研究, 2015, 38(1): 105－113.

[100] Iwata D, Nagahama H, Muto J, et al. Non-parametric detection of atmospheric radon concentration anomalies related to earthquakes[J]. Scientific Reports , 2018, 8(1).

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

北京大学学位论文原创性声明和使用授权说明

原创性声明

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

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

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

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日