# 地震数据分析与预测研究

**作者**：陈笑天，杨妍，林俐，吴金泽，池培德，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：Python数据挖掘、数据预处理、聚类算法、神经网络

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 地震是危害社会和人的生命财产安全的主要自然灾害之一。预测未来可能发生的地震，提前为应对地震做出必要准备，以最大程度减少地震所带来的损失，是人们长久以来追寻的目标。因此，分析过往的地震数据，并从中学习地震的相关规律和制定应对策略是必要的。本案例基于机器学习的主要思想，使用过去几十年的地震信息作为实验数据，运用相关聚类算法分析地震数据，结合神经网络模型对地震震级预测问题进行研究。本案例的核心内容有：（1）下载并使用Scrapy框架爬取地震信息数据，对获取的数据进行预处理；（2）使用Kmeans和DBSCAN聚类算法对比分析地震信息数据，并对分析结果进行可视化；（3）使用神经网络模型预测地震震级。

**关键词**：地震，数据挖掘，聚类算法，神经网络

## **1 引言**

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的主要问题为对地震的历史信息数据进行分析处理，并对地震震级进行预测。基于以上，需引导学生进行的主要内容有：（1）使用Scrapy框架爬取过往的地震数据，运用适当的方法对获取的数据进行清洗和可视化处理；（2）选取适当的聚类算法对预处理后的数据进行聚类分析；（3）选取适当的深度学习模型对地震震级进行预测；

## **2 背景介绍**

地震，作为一种破坏力极强的自然灾害，长期以来危害着社会和人的生命财产安全。提前预测地震，从而为应对地震执行相应的预防措施，最大化的减少地震所带来的损失，是人们一直追寻的目标。随着科技的不断进步，人们对地震的了解正在逐渐深入。然而，即使科技进步的当下，人们尚无法预测地震的到来，并且未来相当长的一段时间内，地震也是无法预测的。尽管如此，近年来，机器学习和深度学习相关算法的提出和发展，很大程度上解决了以前看来难以处理，甚至无法解决的问题。将这些前沿的方法应用于地震的分析和预测是学界的热门领域之一。许多适用于地震学方法的提出，使人类对地震的认识进一步深入。

本案例首先使用基于Scrapy框架的爬虫和预处理的相关技术，获取地震信息数据，并进行可视化等操作。随后，本案例使用Kmeans和DBSCAN算法对处理后的数据进行聚类分析。最后，本案例使用神经网络模型对地震震级进行预测。

本案例以分析地震数据和预测地震震级作为主要研究问题，使用大数据分析与处理、数据挖掘、聚类分析和深度学习的相关方法，具有一定的实际价值，可以满足培养具备实际问题中的思考与解决能力的人才的需求。

## **3 内容**

本案例的主要内容主要分为三个小节，分别为数据获取与预处理、数据聚类分析，以及地震震级预测。

3.1 数据获取与预处理

对于数据获取，本案例所使用数据集的目标数据为1965年至2020年11月4日的。数据集共24252条，由两部分组成，一是和鲸社区的1965-2016全球重大地震数据，共23412条地震数据，网址链接为：<https://www.kesci.com/mw/dataset/58d1172797c4b112cbb826db>；二是利用分布式爬虫来爬取中国地震台网2017-2020/11/04的全球重大地震数据，并存入MongoDB，共840条数据，网站链接为<http://news.ceic.ac.cn/index.html?time=1604666062>。这里使用的分布式爬虫基于Scrapy-Redis框架，该框架在scrapy的架构上增加了Redis，与scrapy相差无几。

数据爬取的运行过程可描述为以下几步。（1）在四台云服务器上部署scrapy爬虫代码。（2）其中master端爬虫运行时把提取到的URL封装成Request存入Redis的eathquake\_scrapy:request数据库中，再从该数据库中Request处理后的网页内容存入Redis的eathquake\_scrapy:items数据库中。（3） 三个slave端从master端的Redis中取出待爬取的Request，下载完网页后把网页内容发给master端的Redis。（4）重复上述3,4步骤，直到master端的Redis的eathquake\_scrapy:request数据库为空，再把master端的Redis的eathquake\_scrapy:items数据库写入MongoDB中。最后从MongoDB中导出CSV格式的数据集，命名为earthquake.csv

对于数据预处理，首先对数据进行清洗。上述得到的数据为[1965-2020全球重大地震数据](https://www.kesci.com/home/dataset/58d1172797c4b112cbb826db/document" \t "_blank)，文件名为earthquake.csv，其中包括27208条地震数据。但是数据中有大部分属性缺失且在本次实验中不会使用，故需要将其手动删除，只保留Date, Time, Latitude, Longitude, Type, Depth, Magnitude这七个属性。其次还需查看数据属性中是否有缺失值，若有，则需要将其删除或替换。

完成数据清洗后，再对数据进行可视化，以得到数据的初步认识，有利于后续对数据的进一步处理。首先分别绘制世界和中国地震的地理分布图，如图3.1和3.2所示。

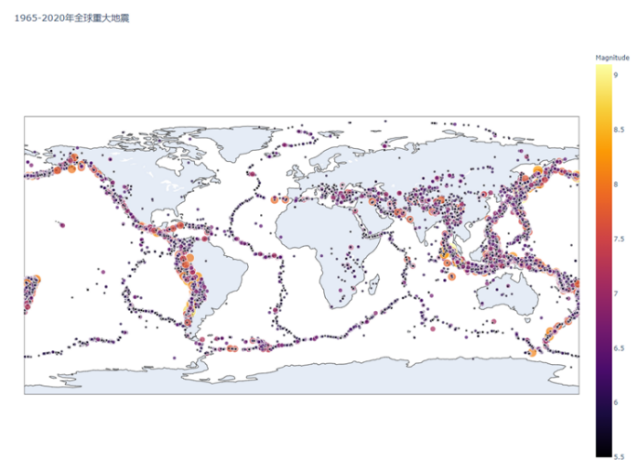


图3.1 世界地震分布图

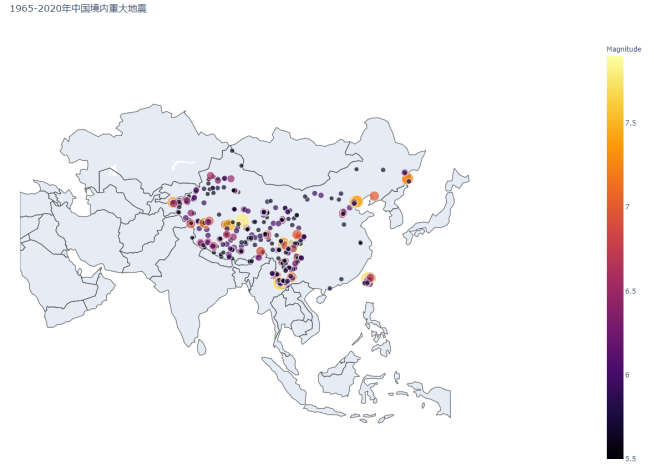


图3.2 中国地震分布图

然后，绘制各省份发生重大地震次数的柱状图和词云图，如图3.3和3.4所示。

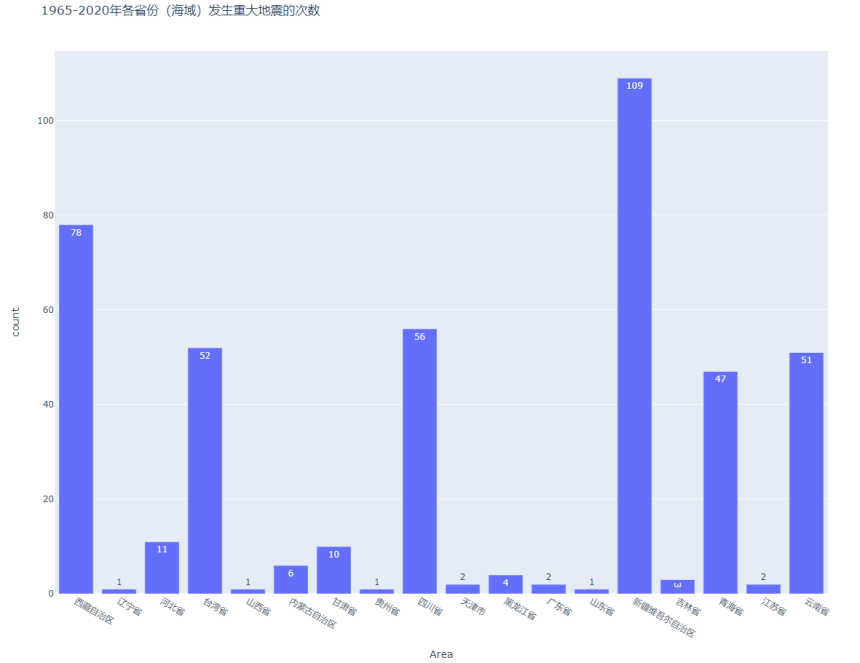


图3.3 各省份发生重大地震次数图



图3.4 各省份发生重大地震次数词云图

最后，绘制世界各地发生地震的震级和震源深度前500名的地理图，如图3.5和3.6所示（更多可视化图像请参阅附件）。

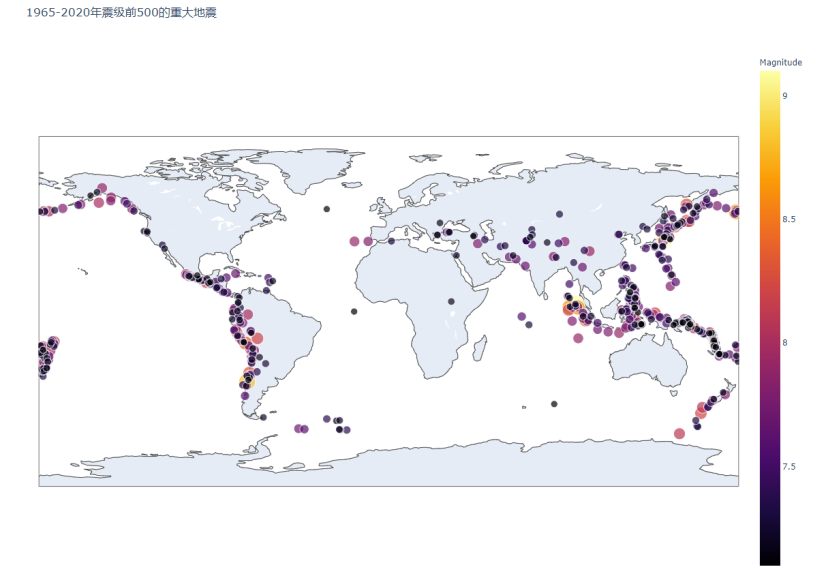


图3.5 世界震级前500名地理图

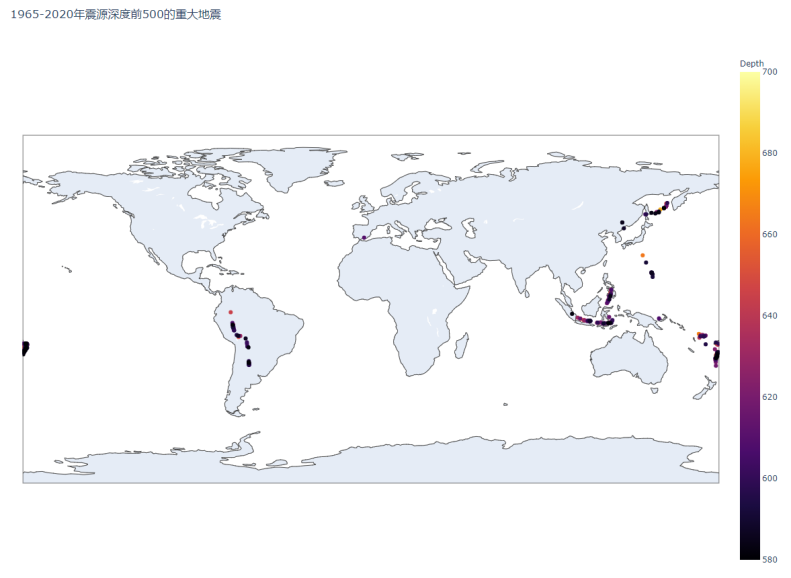


图3.6 世界震源深度前500名地理图

3.2 数据聚类分析

本案例选择聚类算法中的Kmeans算法和DBSCAN算法对爬取的地震数据进行分析。其中，Kmeans算法是最常用的聚类算法，主要思想是:在给定K值和K个初始类簇中心点的情况下，把每个点(亦即数据记录)分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的所有点重新计算该类簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数；DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法是一种基于空间密度的聚类算法，不同于其他传统聚类，其将空间密度较高的区域聚为一簇，生成的簇没有固定形状（与之相比Kmeans聚类形状为圆，密度聚类形状为椭圆）,没有偏移。DBSCAN算法除了可以实现聚类效果，也可以用于寻找数据中的噪声，这是传统聚类算法做不到的。

具体实现上，把1965年到2020年的地震数据作为输入，通过地震的经度纬度坐标来对地震数据进行聚类，用以划分不同的地震带。

对于Kmeans算法，在不同参数下，其得到的聚类结果如图3.7所示。

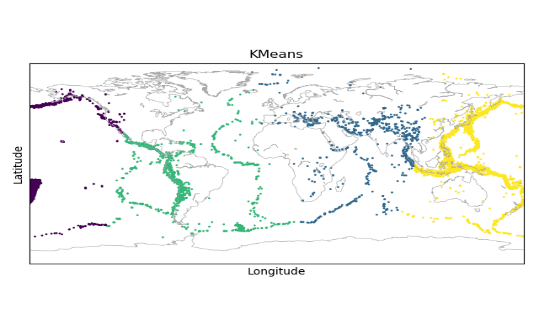
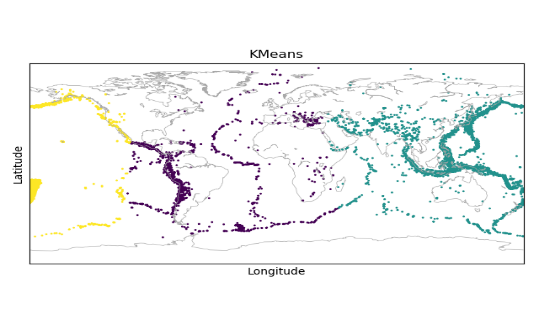
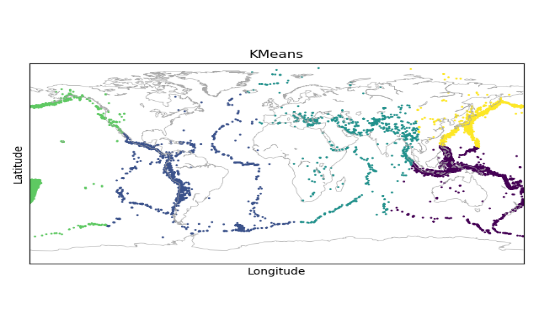
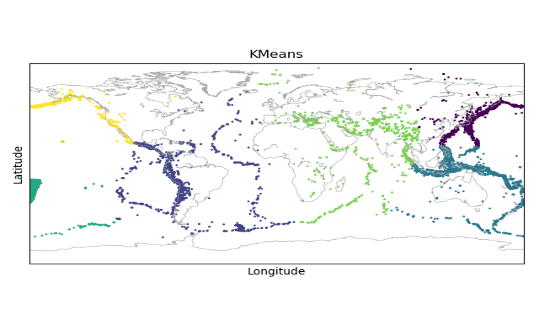


图3.7 Kmeans算法聚类结果图

由图3.7可以看出Kmeans聚类算法虽然能够对地震区域进行聚类，但是其聚类的效果大多呈现球状分布，且某些地震带的聚类结果在板块学说上并不具有相应的说服力。例如，图中的下面两幅子图的日本岛区域，因为Kmeans以球状聚类的缘故导致了原本属于同一个地震带的区域被强行分割开来。

对于DBSCAN算法，在不同参数下，其得到的聚类结果如图3.8所示。

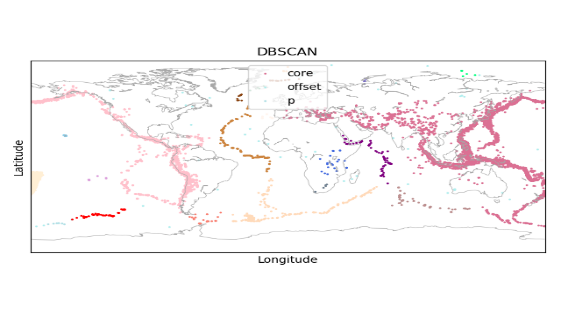
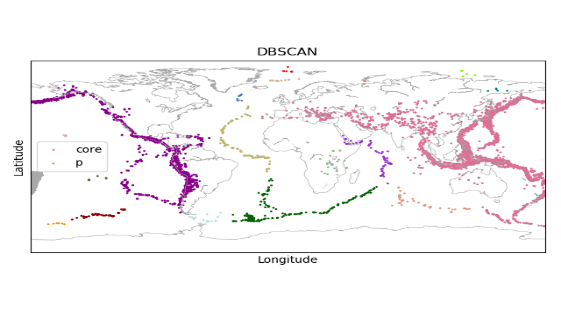
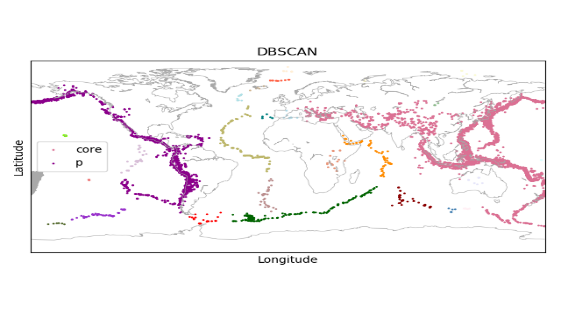
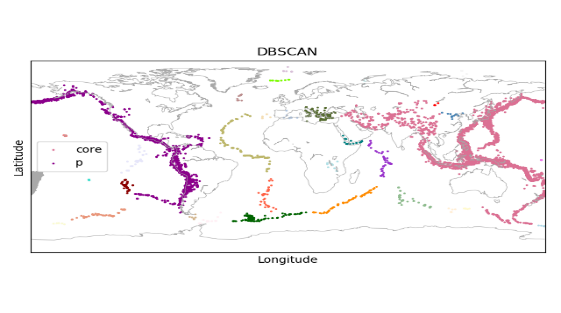


图3.8 DBSCAN算法聚类结果图（MinPts=5，Eps分别为5,6,7,8）

观察图3.8，可以看出DBSCAN算法很好地对地震带进行了聚类，通过调节合理参数，我们可以发现聚类的结果符合地理中的板块学说，且不会出现Kmeans算法中的无缘无故把原本连接的两个板块分开的情况。

下面，将上述两种算法的聚类结果进行对比观察。图3.9和3.10本别绘制了两种算法的聚类结果图与地理学家所绘的地震带图的对比图。

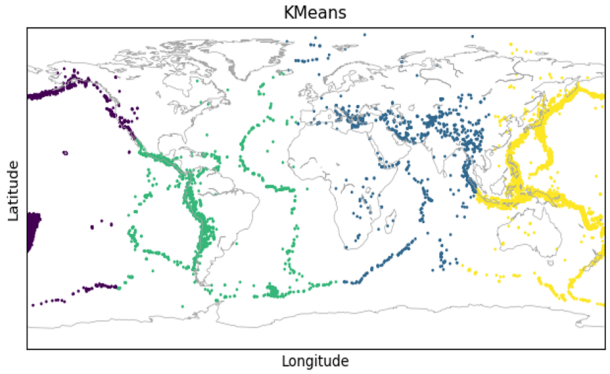
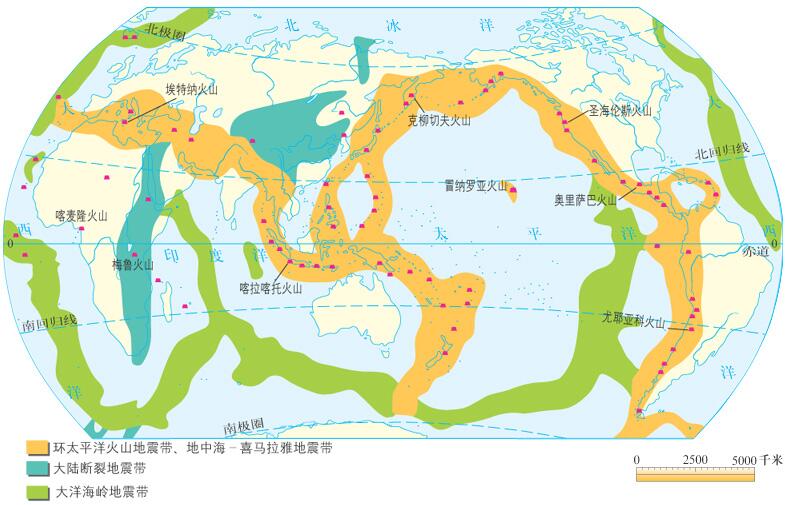


图3.9 Kmeans算法聚类结果与地理学家所绘地震带图的对比图

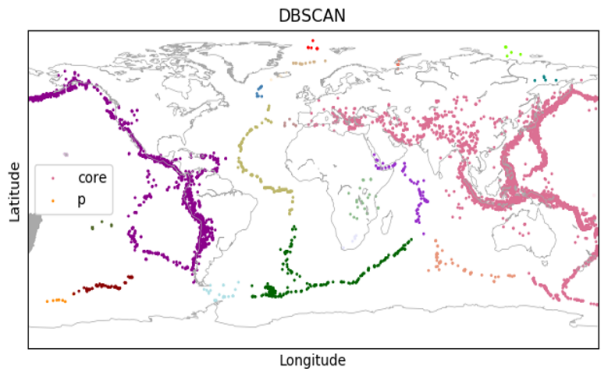
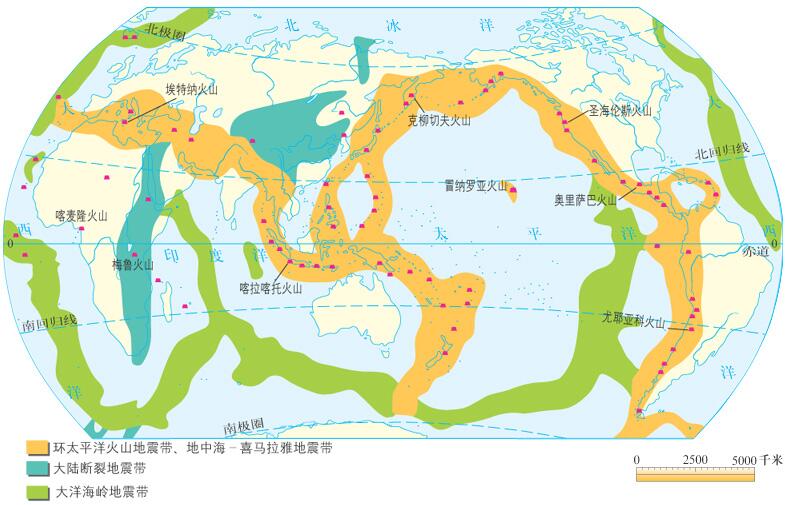


图3.10 DBSCAN算法聚类结果与地理学家所绘地震带图的对比图

观察图3.9和3.10可以看出，Kmeans算法的聚类效果比较生硬，而DBSCAN算法的聚类结果则看上去更自然。同时，对比Kmeans算法的聚类结果图和地理学家所绘的地震带图，可以发现原本在地理学上应该连成块的区域在Kmeans的分类图中则显得非常生硬，好像Kmeans是简单地把地震带分为4个类，在原图上看就貌似一个图的切分。而对于DBSCAN算法，从形式上看该算法的聚类结果与地理学家所绘地震带图非常接近，基本上把地震带给给予标注，而且DBSCAN算法的聚类结果比地理学家手绘的地震带结果更加细节。它在保证地理学模型的基础上显的更加细节，因此，DBSCAN算法聚类的结果图更加具有说服力，利用该算法来对地震带进行聚类是较为合适的。

3.3 地震震级预测

本案例选择使用神经网络模型来进行地震震级预测。对于本问题而言，由于地震震级预测模型的输入是经纬度坐标，输出是未来该地区如果会发生地震则发生地震的震级，所以该问题为一个回归问题，而该模型为一个回归模型。

回归问题也是机器学习三大基本模型中很重要的一环，其功能是建模和分析变量之间的关系。回归问题多用来预测一个具体的数值，如预测房价、未来的天气情况等等。例如我们根据一个地区的若干年的PM2.5数值变化来估计某一天该地区的PM2.5值大小，预测值与当天实际数值大小越接近，回归分析算法的可信度越高。

具体实现上，本案例采用keras框架完成多项式回归模型构建，加入了四个全连接层，每一层采用relu激活函数，使用均方误差作为损失函数，采用Adam优化算法来优化模型。

最终在经度和纬度上的拟合结果如图3.11和3.12所示。



图3.11 经度拟合结果



图3.12 纬度拟合结果

测试结果如图3.13所示。

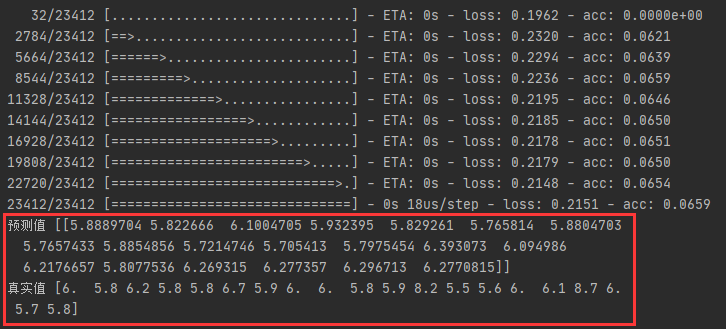


图3.13 测试结果

观察以上三幅图像，可以看出拟合出来的效果欠佳。其主要原因可能在于一下几个方面：（1）数据集的经度和纬度对于震级数据比较离散，且与目标没有较强的关联性；（2）算法没有训练未发生过地震的位置，可能导致了数据集分布不均的问题。后续可以考虑进行更深一步的数据分析，以挖掘出与地震震级更相关的属性，或者设计对于该问题更精准的模型，以提高表现。

## **4 小结**

本案例的主要研究内容是聚类分析地震数据和预测地震震级，研究的目的是掌握数据挖掘和可视化方法和数据的分析和处理方法，同时使用了Python中的scrapy框架来爬取数据，Kmeans和DBSCAN算法聚类分析地震数据，神经网络模型预测地震震级。本案例使用大数据分析与挖掘的相关技术来解决社会实际问题，可以培养学生将理论用于实际的能力。需要指出的是，本案例中所使用的相关技术和算法并非最佳，在教学过程中，可以引导学习发散思维，共同谈论分析该问题的其他方法。

## **附录**

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有scrapy（数据爬取），pandas（数据读取与预处理库），scikit-learn（机器学习算法库），keras（深度学习算法库）。

3. 本案例参考文献如下：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[2] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.