



EL-Picker: 基于集成学习的余震 P 波初动实时拾取方法

申大忠^{1†}, 张琦^{1†}, 徐童^{1*}, 祝恒书^{2*}, 赵雯佳³, 殷子凯¹, 周培伦¹,
房立华⁴, 陈恩红^{1*}, 熊辉^{5*}

1. 中国科学技术大学计算机科学学院, 合肥 230022, 中国

2. 百度在线网络技术(北京)有限公司, 北京 100085, 中国

3. 中国地震局地质研究所, 北京 100029, 中国

4. 中国地震局地球物理研究所, 北京 100081, 中国

5. Rutgers, the State University of New Jersey, Newark NJ 07102, USA

* 通信作者. E-mail: tongxu@ustc.edu.cn, zhuhengshu@gmail.com, cheneh@ustc.edu.cn, hxiong@rutgers.edu

† 同等贡献

收稿日期: 2020-07-15; 接受日期: 2020-07-22; 网络出版日期: 2021-05-08

国家重点研发计划(批准号: 2018YFB1002600)和国家自然科学基金(批准号: 91746301, 71531001, 61703386, U1605251)资助项目

摘要 在实时地震监测中, 地震 P 波(primary wave)的初动拾取任务具有至关重要的作用, 其有助于地震应急响应的及时实施. 虽然此前在该领域已开展了大量的研究, 但是如何从地震分布密集并且充满噪声的监测波形中有效地识别出 P 波仍然是一个具有挑战性的任务. 例如对于大地震的余震监测, 实践中使用的普遍方法仍依赖于专家辅助标注. 本文针对地震实时监测任务, 基于集成学习策略, 提出一个全新的技术框架——EL-Picker, 实现从连续地震波形中自主拾取 P 波的初动到时. 具体而言, EL-Picker 包含 3 个模块, 即触发器、分类器和精化器. 其中, 分类器模块借鉴集成学习策略, 实现对多个个体学习器的整合, 提升整体模型性能. 基于汶川 Ms8.0 地震的余震数据集进行的大量实验, 我们发现 EL-Picker 不仅较好地实现 P 波初动拾取效果, 并且多诊断出 120% 被人工遗漏的地震 P 波. 同时, 实验结果也启发我们探索如何针对不同的地震站台选取个性化的个体学习器构建分类器模块. 此外, 我们进一步地讨论了被人工遗漏的地震波形的规律特点, 用于指导人工地震标注. 这些发现清晰地验证了 EL-Picker 框架的鲁棒性、时效性、灵活性以及稳定性.

关键词 P 波拾取, 机器学习, 集成学习, 汶川余震, 实时地震监测

引用格式: 申大忠, 张琦, 徐童, 等. EL-Picker: 基于集成学习的余震 P 波初动实时拾取方法. 中国科学: 信息科学, 2021, 51: 912-926, doi: 10.1360/SSI-2020-0214
Shen D Z, Zhang Q, Xu T, et al. El-Picker: a machine learning-enhanced robust P-phase picker for real-time seismic monitoring (in Chinese). Sci Sin Inform, 2021, 51: 912-926, doi: 10.1360/SSI-2020-0214

1 介绍

考虑到地震造成的巨大危害,近年来,研究者们已经在实时地震监测领域开展了大量的研究工作.其中一个最重要的任务是地震 P 波初动的拾取,其目的在于鉴别地震纵波(P 波)的确切到达时间.相较于另一种危害性更强的地震波形—横波(S 波),P 波在地壳中传播更快并最先到达地球表面.因此 P 波初动的实时拾取将有助于地震应急方案的及时实施.例如及时启动预警系统、疏散和救援程序等,可极大地减少地震造成的危害.但是如何从地震分布密集并且充满噪声的监测波形中有效地拾取出地震 P 波仍然是一个具有挑战性的问题.例如对于破坏性强的大地震后的连续余震,其往往具有波形频率高、信号规律性弱、幅度范围大以及自然或人为干扰因素多等特点,相应地,余震 P 波初动的拾取也变得尤为困难.

在相关文献中,研究者们已经尝试了多种拾取地震 P 波初动的方法,例如模版匹配^[1~3]以及高阶统计量^[4~6]等.尽管如此,从实际应用的方面考虑,应用最广泛的监测地震的方法仍是基于长短时对比的方法(the ratio of short time average and long time average, STA/LTA)^[7,8]及其相应的变种^[9].其主要思想是评估地震波形的能量变化幅度,当波形能量函数在长时间窗口与短时间窗口中的平均值的比值超过预先设定阈值时则判别当前时间点发生地震事件.但是,由于这些传统的方法往往基于简单的假设,监测能力有限,时常导致误报与漏报问题.因此,为了确保地震监测的正确性以及避免不必要的恐慌,地震监测的实际应用中仍然需要大量的人工标注^[10].

近年来,各种机器学习算法已经在地震监测任务上取得了卓越的效果.其中,常用的机器学习方法包括支持向量机(support vector machine, SVM)^[10]、隐马尔可夫模型(hidden Markov model, HMM)^[11]以及神经网络(neural network, NN)^[12~20].虽然这些机器学习辅助方法有助于减少地震监测中的人工工作量,但是当应用于实时连续的地震波形数据时,机器学习方法的精度、效率以及稳定性难以全面保证.这个缺陷也限制了机器学习方法在实时地震 P 波初动拾取任务中的实践.具体而言,文献中相关研究主要集中于两个方面:或者从候选时间窗口样本中鉴别包含地震的样本^[14,16,18],或者在包含地震的窗口中拾取精确到达时间^[15,17].实际上,大部分研究工作是基于滑动窗口的策略,其往往需要大量的计算资源,因此很难在实时连续地震监测系统上部署.很少有研究者考虑从实时、连续的地震波形数据中选择合适的时间窗口作为机器学习模型的输入候选集,以此减少计算资源需求.

本文提出了一种基于集成学习策略的机器学习框架 EL-Picker,用于自主地震 P 波初动实时拾取.其基本的思想在于先使用需求更少计算资源的传统地震学方法,例如 STA/LTA,过滤并剔除大部分噪声信号.随后运用机器学习方法从保留下来的信号中判别真实的地震 P 波.具体而言,如图 1 所示,EL-Picker 由 3 个模块组成,即触发器、分类器和精化器模块.触发器模块使用传统的地震学方法检测潜在的地震 P 波初动,并滤除大量的噪声信号.分类器模块引入有效的机器学习模型评估每个潜在的地震 P 波到时的可信度.最后,精化器模块从高置信度的 P 波候选集中识别最准确的 P 波初动时间点,并剔除异常信号点.值得一提的是,我们使用集成学习策略具体实现分类器模块.该策略的有效性已经在各种实践应用以及数据挖掘竞赛中得到了验证^[21].在实验部分,我们以汶川 Ms8.0 地震的余震数据集为例,利用多个地震台站记录的连续波形数据模拟实时地震监测场景,进行了大量的实验.实验结果以及进一步的讨论分析清晰地展示了我们框架的鲁棒性、时效性、灵活性与稳定性.此前,在 2017 年举办的余震捕捉 AI 大赛^[22]中,我们借助于 EL-Picker 的初步版本,从全世界 1143 个队伍中取得初赛冠军和决赛亚军的好成绩.

综上所述,本文的主要贡献分为以下 3 个方面.

- 结合传统地震学方法以及机器学习方法提出了一个由 3 个模块构成的地震 P 波初动拾取框

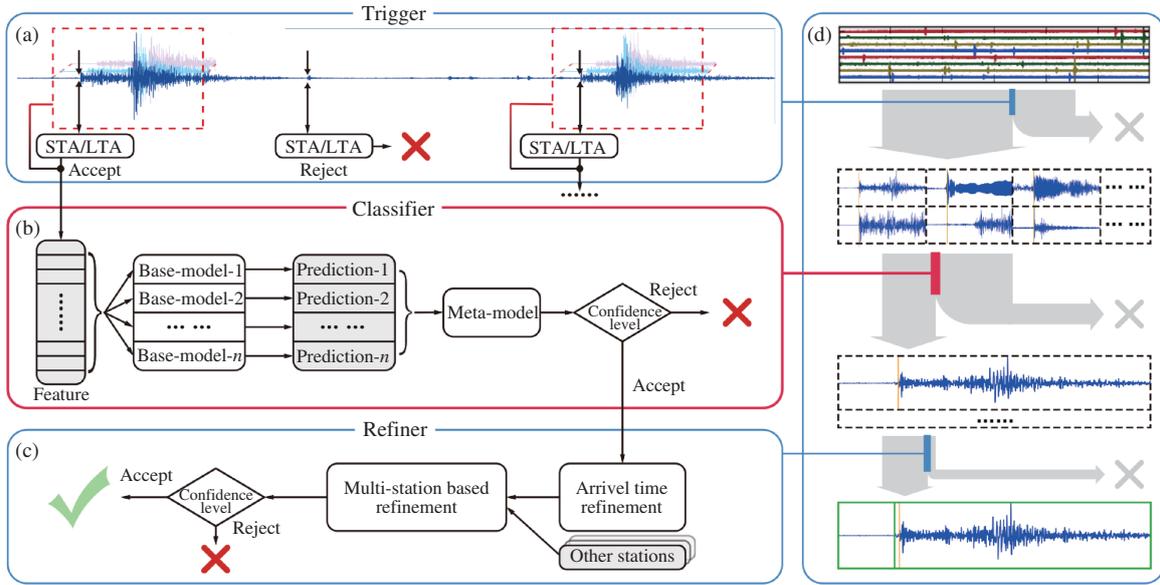


图 1 (网络版彩图) EL-Picker 框架图

Figure 1 (Color online) An overview of the EL-Picker framework

架——EL-Picker, 实现了从连续地震波形中自主、实时地进行 P 波初动拾取。

- 基于集成学习策略设计了一个具备高鲁棒性的机器学习模块用于地震 P 波判别。
- 大量的实验与讨论分析清晰地展示了 EL-Picker 框架的鲁棒性、时效性、灵活性与稳定性。

2 相关工作

本文的相关工作可分为两大类: 基于传统地震学的方法和基于机器学习的方法。

2.1 基于传统地震学的方法

作为一个有利于社会公益的问题, 地震波形初动拾取已经在地震学领域引起了广泛的研究关注。如今, 研究者们已经从不同的角度提出了多种有效的解决方案。例如模版匹配, 其主要思想认为发生机制类似的地震之间产生波形数据也是相似的。Gibbons 等^[1]是最早将模板匹配应用于地震学领域的研究者之一。此后, 大量的拓展方法被相继提出, 如自相关^[23,24]和相似性搜索^[25]等^[2,3]。虽然这些方法均被证明是寻找相似地震的有效解决方案, 但是由于巨大的计算复杂度, 这些方法很难被应用于实时的地震波形初动拾取任务。此外, 几种统计方法也被用来处理地震波形数据。例如高阶统计量 (higher order statistical function, HOS)。Saragiotis 等^[4,5]是利用 HOS 函数鉴别 P 波初动的先驱之一, 其将偏度和峰度函数引入到波形拾取中。随后, Küperkoch 等^[6]进一步改进了 Saragiotis 方法, 通过加权平均的方式实现更好的波形拾取效果。然而, 考虑到地震监测中的时效性需求, 在具体实践中应用最广泛的方案还是基于 STA/LTA 方法及其变种, 如 Allen Picker^[8], BK87^[26]等^[9,27]。其中, 文献^[9]中提出的 FilterPicker 算法表现出对地震事件更高的敏感性, 具备良好的鲁棒性与高效性。

2.2 基于机器学习的方法

近年来, 机器学习方法已经被成功应用于地震学的许多领域, 如地震定位^[28]、地震震级判别^[29]、

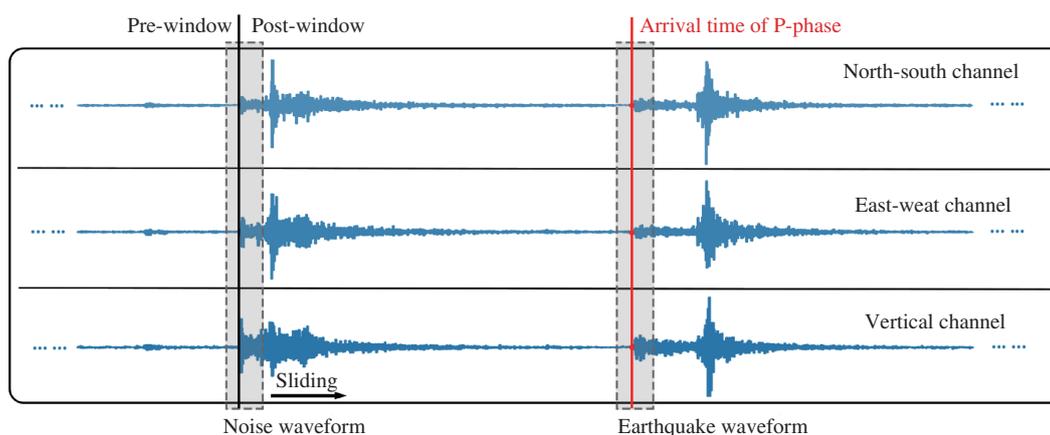


图 2 (网络版彩图) 3 个空间维度上的真实地震波形数据示例

Figure 2 (Color online) An example for the real-world seismic waveforms on three dimensions

地震分类^[30]和烈度估计^[31]。值得一提的是,针对地震监测和地震波形的初动拾取任务,多种机器学习方法的有效性已经被充分地探索与讨论。例如, Ruano 等^[10]设计了一种基于支持向量机(SVM)和多层感知器(multi-layer perceptions, MLPs)的地震监测系统。Beyreuther 等^[11]应用隐马尔可夫模型(HMMs)对中小地震进行检测和震源距离相关的分类。Wang 等^[12]提出将全连接神经网络应用于地震事件检测。此后,神经网络因其强大的灵活性和泛化能力在地震监测领域受到了广泛的研究关注^[13~15]。最近,除了传统的全连接网络外,卷积神经网络也被用来提高地震监测性能。Perol 等^[16]设计了一种基于时间窗口的地震监测卷积神经网络(conventional neural network, CNN)结构。Zhu 等^[17]基于 CNN 设计模型用于判别每个信号点作为波形初动的概率。Zhang 等^[18]提出的 CNN 结构通过同时捕捉长期与短期的地震特征实现更好的地震鉴别效果。此外,由于声音信号和地震信号的相似性,研究人员也尝试探索语音识别中得到广泛应用的长短期记忆网络(long short-term memory, LSTM)在地震监测中的效果^[32]。例如, Mousavi 等^[19]以及 Wiszniowski 等^[20]分别设计了基于 LSTM 的地震监测系统,并讨论与展示了 LSTM 在微震监测中的优势。然而,与传统的 STA/LTA 方法相比,基于机器学习的地震监测方法往往需要更大的计算资源,并且现有的工作主要集中于两个方面:或者从候选时间窗口样本中判别包含地震的时间窗口^[14,16,18],或者在已知包含地震的窗口中拾取地震精确到^[15,17]。因此,在不进行候选集初步筛选的情况下,针对每个时间窗口或每个信号点进行逐一判别是不现实的,这妨碍了复杂的机器学习模型在连续地震实时监测系统部署。

3 EL-Picker 框架

在地震波形数据中,我们在 3 个空间维度分量中记录地震造成的震动波形,即垂直方向 Z、南北方向 N,以及东西方向 E。图 2 中展示了这 3 个维度中地震波形样例(图 S1 展示了更多的例子)。P 波总是先到达地球表面并在 3 个维度中引起波形的振幅波动,地震学领域通常把这些振幅波动的起始点(用红色实线标记)称为 P 波初动,表示一次地震事件的起始时间。此外,我们也注意到地震波形中还存在一些具有误导性的噪声波形(如,黑色线标记处)。直观来看,针对 P 波初动的鉴别,我们需要对比其周围的短时间窗内的波形变化。沿着这种思路,P 波初动拾取的目标可以被定义为:给定 3 个空间维度上的地震波形数据,我们希望根据每个信号点周围的短时间窗口,从各种噪声信号中自动提取

每个地震事件的 P 波初动. 为此, 我们在本节中提出 EL-Picker 框架用于连续地震波形中的实时 P 波初动拾取. 如图 1 所示, EL-Picker 由 3 个模块构成, 即触发器、分类器以及精化器模块. 接下来, 我们将分别介绍它们的技术细节.

3.1 触发器模块

在触发器中, 我们使用传统地震学方法, 例如 STA/LTA 及其变种, 在每个地震台站收集的波形数据上进行噪声信号过滤以及潜在 P 波拾取. 具体而言, 当信号点周围波形能量函数上的波动大于预先设定的阈值时, 其被判别为潜在的 P 波初动. 同时为了捕捉尽可能多的潜在 P 波, 我们设置较小的阈值. 本文借鉴一个 STA/LTA 的变种, 即 FilterPicker^[9] 模型, 来具体实现触发器模块. 首先, 我们对每个空间维度上的波形数据分别进行带宽为 (2.5~5 Hz, 5~10 Hz, 以及 10~20 Hz) 滤波处理. 随后, 我们借鉴文献 [9] 中定义的能量函数评估每个点上的信号波动大小. FilterPicker 模型包含多个需要预先设定的参数, 即触发阈值 S_1 、时间宽度 T_{up} 以及平均阈值 S_2 . 在某时刻 t , 如果能量函数值超过 S_1 , 并且在随后 T_{up} 长度的时间窗口内的能量函数平均值超过 S_2 , 则该时刻 t 将被记录为潜在的 P 波初动.

3.2 分类器模块

由于触发器模块中的阈值设置相对较小, 拾取的潜在 P 波初动中误报的比例相应较高. 因此, 我们必须进一步筛选出被误判的候选信号点, 并以较高的置信度和可靠性判别真实的地震 P 波. 为此, 我们设计了分类器模块对触发器提取的潜在 P 波进行再次判别, 具体而言, 通过在初动时间点附近的固定长度的时间窗口提取的特征, 我们训练分类器模型用于区分真实地震 P 波与其他信号 (如噪声). 其中, 我们利用集成学习策略^[33] 对分类器模块进行模块化设计. 集成学习策略的主要思想在于使用元学习器整合多个个体学习器的输出, 提升模型整体性能. 一般来说, 个体学习的数量越多, 集成学习的性能越鲁棒^[34]. 在 EL-Picker 中, 我们选择 9 个被广泛使用的机器学习分类模型作为个体学习器, 即线性支持向量机 (SVM-linear)^[35]、多项式支持向量机 (SVM-poly)^[10]、信息增益决策树 (Tree-gini)、熵增益决策树 (Tree-entropy)^[36]、K 近邻 (K-nearest neighbors (KNN))、随机森林 (RandomForest)^[37]、Adaboost^[38]、Logistics 回归 (LogisticRegression)^[39] 和高斯贝叶斯 (GaussianNaiveBayes)^[40]; 并使用线性模型, 例如 Logistics 回归, 加权整合所有个体学习器的判别得分, 提升整体模块的性能. 下面展示了分类器模块整体的训练与验证过程.

- 对训练数据集进行 5 等分, 并分别对于每个个体学习器在 5 个子集进行交叉训练与验证^[41]. 具体而言, 我们依次将 1 个数据子集作为测试数据, 而其他 4 个数据子集作为训练数据. 由此, 每个子集中的每个样本都会被作为测试数据, 并得到每个个体学习器预测的得分.
- 使用每个样本得到的 9 个个体学习器预测的得分作为输入训练元学习器 Logistics 回归模型.
- 在全量训练集上分别训练 9 个个体学习器, 并结合上一步学习到的元学习器的参数, 对个体学习器进行线性加权, 得到完整的分类器模块.
- 在验证阶段, 对于每个时间窗口, 利用上一步得到的分类器模块可以判别其为地震 P 波的置信得分. 置信得分超过设定阈值的时间窗口被认为是高置信度 P 波初动.

3.3 精化器模块

由于 STA/LTA 算法通常无法灵敏地识别地震 P 波初动的准确到达时间, 因此, 我们希望进一步地提升地震 P 波初动拾取的精度. 同时, 注意到同一地震事件往往会被多个相邻的站台监测到. 直观来看, 同时结合多个站台的波形信号有助于减少单个站台数据中存在的误差. 依据这些观察, 我们进

一步设计精化器模块用于过滤掉异常样本并筛选出最准确的地震 P 波初动. 具体而言, 我们首先使用在地震学领域被广泛认可的 AIC^[42~45] 模型去精化每个候选地震 P 波的到达时间. 随后, 对于每一个被识别出的地震 P 波初动, 我们使用监测出同一地震的台站个数作为附加的置信水平评估指标, 进一步过滤异常样本. 总的来说, 本模块主要由两部分构成, 其执行步骤如下:

- 使用 AIC 模型在每个 P 波初动的判别时间点附近 ± 1 s 的短窗口内搜索精确的到达时间.
- 对于任两个不同站台监测到的 P 波初动, 我们认为当两个 P 波初动的时间差小于 P 波在两站台间的传播时间 D/v_p 时, 两个 P 波对应于同源地震. 其中, P 波传播速度 v_p 被设置为小于四川省的平均值, 即 $v_p = 5.5$ km/s; D 为两个地震监测站台的距离. 所有只被一个站台记录的 P 波到时被认为是异常样本而被剔除.

4 数据描述

本文使用两个余震数据集, 即汶川 Ms8.0 地震的余震数据集 (中国四川, 2008.05.12), 以及 Mw 6.0 South Napa 地震的余震数据集 (美国加利福尼亚, 2014.08.24). 数据全部由中国地震局地球物理研究所中国地震台网数据管理中心提供. 具体而言, 每一天的余震波形数据包含 Z, E 和 N 三个维度方向上的波形变化记录, 采样率为 100 Hz. 同时, 为了确保数据的质量, 我们进一步去除包含异常片段的样本, 例如由监测仪器异常导致的噪声. 最终, 汶川余震数据集涉及连续 4 周内的 12696 个地震记录, 每周依次包含 3447, 3442, 3997 和 2810 个 P 波初动记录. Napa 余震数据集包含 680 个 P 波初动记录. 随后, 为了训练分类器模块, 我们从触发器模块保留的潜在 P 波初动中选取了 5 倍于真实 P 波初动的负样本. 具体而言, 我们认为与所有真实标注 P 波初动的时间误差都超过 0.4 s 阈值的信号时间点为负样本. 这些负样本包含但不限于噪声, 不精确的 P 波初动, 以及 S 波初动.

接下来介绍特征提取过程. 具体而言, 为了分类器模块的训练与应用, 我们需要针对每一个有待判别的信号点临近区域的时间窗口进行特征抽取. 值得注意的是, 时间窗口的长度是可以进行调整设定的. 这里我们设置位于信号点之前的前置窗口长度为 5 s. 而位于信号点之后的后置窗口长度可以在 5 s 到 20 s 之间进行调整. 相应地, 我们可以从不同时间窗口中抽取 679~715 维特征向量. 表 1 展示了所有特征的计算过程, 其中子时间窗口使用开区间表示, 潜在 P 波初动信号时间点被标记为 0. 正负数分别代表着前置与后置窗口的长度. 为了方便介绍, 我们大致划分这些特征为 4 组.

- **振幅波动特征.** 考虑到 P 波初动附近会出现剧烈的振幅波动, 对于不同的后置窗口长度, 我们选取 5~8 个子时间窗口, 然后对于每个子窗口中 3 个空间维度上的波形记录分别进行 2~10 Hz 和 10~20 Hz 的带宽滤波, 并计算滤波后波形振幅的平均值与方差. 该过程总共提取出 60~96 维振幅波动相关特征.

- **最大振幅特征.** 考虑到 S 波总是紧随 P 波到达地球表面, 并导致波形数据上, 尤其是在水平方向上 (E 和 N), 的极大波动. 由此, 我们首先对数据进行 2~10 Hz 和 10~20 Hz 的带宽滤波, 然后分别提取 3 个空间维度上波形振幅极大的位置. 随后在 N 与 E 两个维度上的波形, 计算极大值点附近 1 s 邻域内的振幅平均值与方差. 该过程总共提取出 14 维最大振幅相关特征.

- **频谱特征.** 考虑到波形频谱的重要意义, 我们划分信号点附近 1 s 邻域为 10 个片段, 对于 3 个空间维度上的每一个片段, 我们使用 0~50 Hz 之间的 9 个相邻带宽的滤波器进行滤波, 并计算片段内振幅平均值与方差. 该过程总共提取出 540 维特征.

- **其他特征.** 一些其他的特征用于进一步提升分类器模块性能, 例如振幅比率、平均差、极化斜率和包络斜率 (具体的计算过程参见表 S3). 该过程总共提取出 65 维特征.

表 1 特征提取的计算过程
Table 1 Detail computation procedures for classification features

	Time window (s)	Bandwidth (Hz)	Channel	Computation process for each time window x with each bandwidth in each channel	Number of features					
Amplitude fluctuation	$(-5, 0), (0, AN), (-1, 0), (0, -1)$ $(5(i-1), 5i)$ $1 \leq i \leq \text{int}(AN/5)$	2~10 10~20	E	mean(x) var(x)	48 + $12 \times \text{int}(AN/5)$					
			N							
			Z							
Maximal amplitude	(2, AN)	2~10 10~20	E	idx = $x.\text{index}(\text{Max}(x))$	12					
			N	mean(x [idx - 1, idx + 1]) var(x [idx - 1, idx + 1])						
			Z	idx = $x.\text{index}(\text{Max}(x))$						
Spectral waterfall	$(-0.2, 0), (0, 0.2)$ $(-0.4, 0), (0, 0.4)$ $(-0.6, 0), (0, 0.6)$ $(-0.8, 0), (0, 0.8)$ $(-1.0, 0), (0, 1.0)$	0.5~0.833 0.833~1.389 1.389~2.314 2.314~3.858 3.858~6.430 6.430~10.717 10.717~17.816 17.816~29.768 20.768~49.615	E	mean(x) var(x)	540					
			N							
			Z							
			Other features			(-5, 5)	1.389~2.314 2.314~3.858 3.858~6.430 6.430~10.717 10.717~17.816	E	RMS amplitude ratio($x[0, 5], x$)	60
								N	Mean difference($x[0, 5], x$)	
								Z	Envelope slope (x)	
				Polarization slope($x[-5, 5]$ in channel E, N, Z)	5					

5 实验结果

为了全面地评估 EL-Picker 的性能, 我们在汶川余震数据集上开展 4 折交叉验证的实验流程. 具体而言, 对于 4 周数据, 我们依次使用 1 周的数据作为测试数据, 而另外 3 周的数据作为训练数据进行模型训练与测试. 此外, 为了进一步验证我们框架的鲁棒性, 我们增加了一个额外的测试实验, 使用汶川数据集整体作为训练数据, 在 Napa 数据集上进行测试验证. 实验中, 我们设置触发器模块中参数为 $S_1 = 6$, $S_2 = 2$ 和 $T_{\text{up}} = 0.3$ s; 设置分类器的判别阈值为 0.5. 此外, 我们使用精确度 (precision)、召回率 (recall) 以及 F 值 (F-score) 作为性能评估指标.

首先, 我们对分类器模块的性能进行评估. 具体而言, 对于用于训练与测试的正负样本, 我们将专家人工标注的 P 波初动邻域的时间窗口作为正例样本. 相应地, 负样本从触发器模块保留的潜在 P 波初动中选取. 其中, 后置窗口的长度被设置为 20 s. 对于作为性能对比的基线方法, 我们选用了另外 3 种机器学习方法, 即用于时间序列信号分类的神经网络 LSTM^[46]、卷积神经网络分类方法 Inception^[47] 以及针对地震监测的 CNN 模型 ConvNetQuake^[16]. 表 2 的左半部分总结了分类器模块以及基线的平均性能, 从实验结果中, 我们发现分类器模块能够有效地判别与区分真假 P 波, 并且在 F 值评估指标上始终优于其他基线. 更有趣的是, 我们的分类器模块在 Napa 数据集也表现出相对稳定的性能, 而其他基线的性能却有很大程度的下降, 特别是对于召回指标而言. 这个现象清晰地证实了分类器模块的鲁棒性. 同时, 我们也发现, 无论分类器模块还是基线方法在 Napa 数据集的性能都不如在汶川数据集中的表现. 这可能是由于汶川和 South Napa 两地之间巨大的地理位置与地质环境的

表 2 分类器模块与个体分类器的性能评估结果

Table 2 The performance of the Classifier module and the base models

Methods	Precision	Recall	F-Score	Methods	Precision	Recall	F-Score
LSTM	0.7918	0.7494	0.7682	SVM-linear	0.8832	0.8932	0.8878
Inception	0.8122	0.8144	0.8124	LogisticRegression	0.8822	0.8939	0.8876
ConvNetQuake	0.8808	0.8346	0.8563	Adaboost	0.8861	0.8869	0.8860
Classifier	0.9027	0.8869	0.8941	RandomForest	0.8816	0.8828	0.8814
Testing on Napa data set				SVM-poly	0.8703	0.8101	0.8384
LSTM(N)	0.1511	0.4328	0.2240	Tree-entropy	0.8336	0.8319	0.8321
Inception(N)	0.9708	0.1976	0.3284	Tree-gini	0.8237	0.8266	0.8246
ConvNetQuake(N)	0.9848	0.1940	0.3242	KNN	0.7839	0.7831	0.7824
Classifier(N)	0.8059	0.9226	0.8603	GaussianNaiveBayes	0.6956	0.8876	0.7789

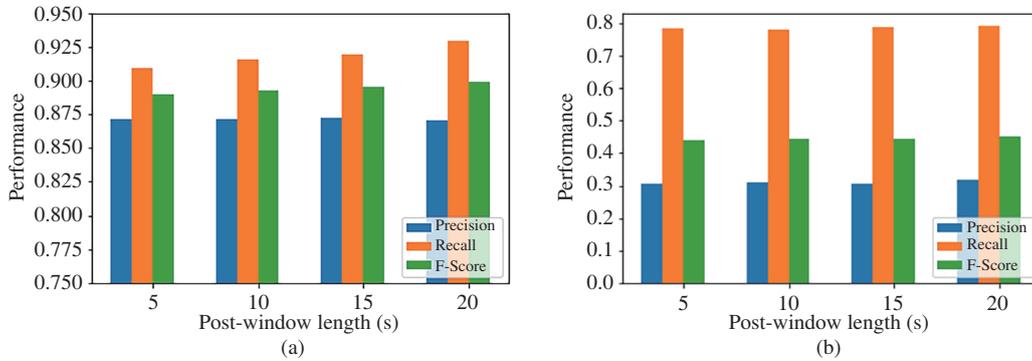


图 3 (网络版彩图) 不同时间窗口下的分类器模块 (a) 与 EL-Picker 框架 (b) 性能评估结果

Figure 3 (Color online) Sensitivity to the time window length within Classifier module (a) and EL-Picker framework (b)

差异导致了两地之间的地震波形模式也具有很大的差异. 此外, 得益于集成学习策略, 我们的分类器模块的性能也优于每个个体学习器. 如表 2 的右半部分所示, 尽管基于我们精心设计的特征个体分类器也实现了具有竞争力的性能, 但在精确度和 F 值评估指标下, 分类器模块的整体性能要优于所有个体学习器.

随后, 我们验证分类器模块性能针对时间窗口长度的敏感性. 直观来看, 较大的时间窗口可以提供更多的波形信息, 提高检测性能, 而较短的时间窗口可以节省监测时间, 提升地震监测的时效性, 有利于应急方案的实施. 因此, 在鲁棒性与时效性之间如何寻求平衡是一个值得讨论的话题. 为此, 我们针对不同长度的后置时间窗口 (即 5, 10, 15 和 20 s) 重复上述分类器模块交叉验证试验. 根据图 3(a) 中的结果, 我们发现在不同的时间窗口下分类器模块的性能相对稳定.

接下来, 我们评估整体 EL-Picker 框架的性能. 我们在实验中使用完整且连续的地震波形数据模拟一个真实的应用场景. 如果一个潜在的 P 波初动和专家标注的 P 波初动之间的时间差小于 0.4 s, 我们认为潜在的 P 波初动为“正确”的. 通过这种方式, 我们针对 EL-Picker 的完整框架 (即, 触发器 + 分类器 + 精化器) 重复了上述两个交叉验证实验. 并通过去除分类器模块的方式构造一个对比基线, 即, 触发器 + 精化器. 事实上, 这个基线可以被看作是传统地震监测方法的广义版本. 表 3 展示了本实验的结果, 其中“CV_n”代表第 n 组交叉验证实验, “CV-Napa”代表在 Napa 数据集上进行的测

表 3 连续地震波形数据下的 EL-Picker 框架性能评估结果
 Table 3 The performance of the EL-Picker framework with continuous seismic waveforms

	Trigger + Refiner			EL-Picker		
	Precision	Recall	F-Score	Precision	Recall	F-Score
CV1	0.0017	0.9005	0.0034	0.3165	0.7926	0.4524
CV2	0.0019	0.9102	0.0039	0.4163	0.7435	0.5337
CV3	0.0018	0.9266	0.0036	0.4349	0.7317	0.5456
CV4	0.0017	0.9221	0.0034	0.4167	0.7338	0.5316
CV-Napa	0.0013	0.9147	0.0026	0.4800	0.4412	0.4598

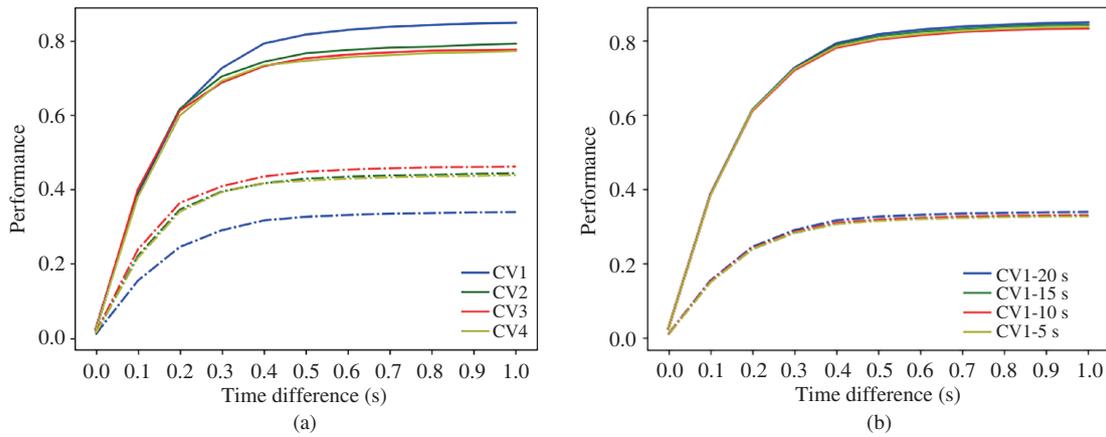


图 4 (网络版彩图) 不同时间误差阈值下的 EL-Picker 框架性能评估结果. (a) 不同交叉验证实验下的精确度 (实线) 与召回率 (虚线) 曲线. (b) 不同后置时间窗口下的精确度 (实线) 与召回率 (虚线) 曲线

Figure 4 (Color online) The sensitivity to the time difference within the EL-Picker framework (a) The precision (solid) and recall (dashed) curves for the validations of EL-Picker with different time difference boundaries. (b) The precision (solid) and recall (dashed) curves for the validations of EL-Picker with different time difference boundaries and different post-window lengths

试实验. 我们发现 EL-Picker 在精确度评估指标上明显优于基线, 且大约有 200 倍的提升. 尽管分类器模块在召回率指标上逊于基线, 但通过综合考虑精确性和召回率 (即通过考虑 F 值指标), EL-Picker 框架的性能得到了充分的体现. 此外, 我们还进一步评估我们的框架对分类器模块中选取的时间窗口长度的敏感性. 如图 3(b) 所示, 实验结果证实了我们的框架在不同的时间窗口长度下具有稳定的性能.

此外, 我们发现一些被 EL-Picker 以高置信度捕捉的 P 波初动并没有被专家标记. 这种现象有两种可能的解释. 一方面, 在连续地震波形中, EL-Picker 有时会捕获到一个潜在 P 波到时, 但它的误差超出了预先定义的 0.4 s 阈值; 考虑到这种情况, 连续地震波形的实时监测要比在给定时间窗口中进行地震判别的分类任务困难得多. 另一方面, 我们忽略了手动标记的数据集可能包含微小的遗漏. 这个问题将在本文后面讨论.

然后, 我们进一步评估 EL-Picker 对判别时间误差阈值的敏感性. 为此, 我们在不同判别时间误差阈值下, 评估框架的性能. 如图 4 所示. 精确度和召回率指标都随着阈值的增加而增加, 并且在 0.4 s 后趋于收敛, 这意味着识别的 P 波到时和专家标记到时之间的时间差大部分小于 0.4 s.

最后, 我们分析 EL-Picker 整体框架在实践中的运行时间. 具体而言, 我们选择了一个地震监测台

表 4 EL-Picker 框架中每个模块的运行时间分析

Table 4 The time consumption of each module in the EL-Picker framework

	Trigger	Classifier	Refiner
Number of signals	60480000	479204	12033
Average running time (ms)	0.0025	706.3258	16.8616

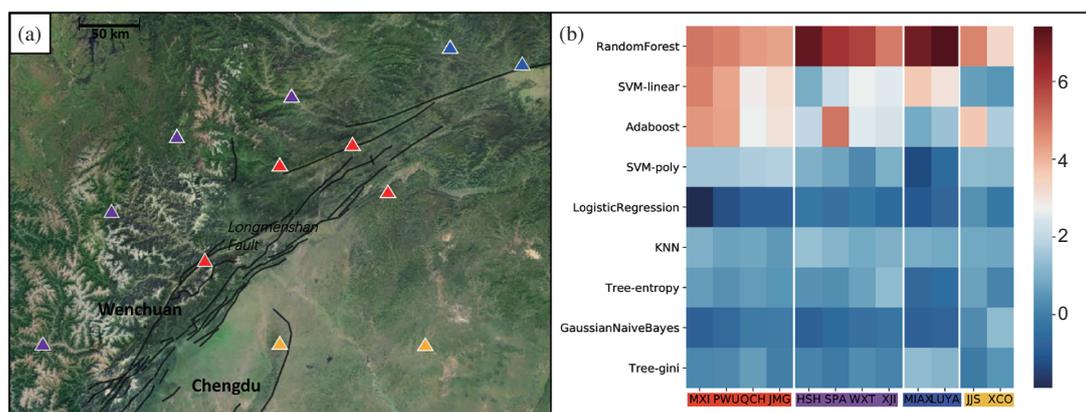


图 5 (网络版彩图) 12 个台站的聚类结果

Figure 5 (Color online) The clustering results of 12 monitoring stations

站收集的第一周波形数据作为测试数据, 并在一个单核 Intel Xeon E5 2.2 GHz 中央处理器 (CPU) 上运行我们的框架来分析模型运行时间. 表 4 显示了 EL-Picker 框架中每个模块针对每天数据处理的平均信号数量和每个信号点对应的平均运行时间. 结果表明, 基于 STA/LTA 模型的触发器模块具有较低的时间复杂度, 能够在短时间内滤除大部分噪声, 其中只有少于 0.08% 的信号点被保留, 减少分类器模块中不必要的计算资源浪费. 此外, 分类器模块能够进一步去除 97.5% 的虚假 P 波初动, 提高精化器模块的计算资源利用率. 同时, 分类器模块中特征提取部分以及集成学习策略, 都可通过并行化策略提高时间效率. 例如并行化提取不同类别的特征以及并行化运行不同个体学习器. 另外, 一些个体学习器同样是基于集成学习策略构建的, 例如 RandomForest 和 Adaboost, 可以通过并行化进行加速运行. 实验证明, 与不使用并行策略相比, 通过这些简单的并行化策略, 平均运行时间可以减少至原本运行时间的 17.2%. 更多的讨论参见补充材料.

基于以上结果, EL-Picker 框架的鲁棒性和时效性已经得到了有效的验证.

6 讨论分析

为了确保 EL-Picker 框架在实际场景中的适用性, 我们进一步探索分析 EL-Picker 框架的一些相关特性, 主要包含两个部分, 即台站聚类分析与遗漏 P 波规律分析.

台站聚类分析. 实际上, 由于不同地理环境的差异, 即使对于同一次地震, 不同台站记录的地震波形可能具有不同的特点. 由此, 我们首先探讨 EL-Picker 框架是否能够调整性地应用于各种台站收集来的地震波形. 具体来说, 对于汶川数据集中的每一个台站, 我们使用 4 周全部数据训练分类器模块, 然后把集成学习模块训练得到的 9 个个体学习器的权重参数, 作为台站的个性化表征向量 (表 S4 展示了具体的权重参数). 随后, 我们利用 Kmeans 聚类方法^[48] 划分 12 个台站为 4 类. 如图 5 所示,

我们在地图中展示了 4 类台站的地理位置分布和不同台站对应的个体学习器的权重参数分布. 我们发现红色簇的台站坐落在龙门山断裂带附近, 紫色簇的台站环绕在山脉周边, 蓝色簇的台站坐落在山脚, 而黄色簇的台站远离断裂带, 坐落在平原上. 我们可以发现聚类结果以及个体学习器的权重参数与地质环境高度相关. 例如, 对于断裂层附近的台站 (即红色聚类), RandomForest^[37], SVM-linear 和 Adaboost^[38] 的权重明显高于其他基础模型. 对于整个山区的周围地区 (即紫色簇和蓝色簇), RandomForest 是权重最高的, 而对于平原上的台站 (即黄色簇), 个体学习器的权重差异较小.

遗漏 P 波规律分析. 在第 5 节, 我们观察到的一个特殊问题, 即 EL-Picker 会以高置信度捕捉到一些没有被专家标注的地震 P 波初动. 为了合理地验证我们框架的性能, 我们必须校验框架设计中是否存在缺陷, 或者专家手动标记的数据集是否包含微小的遗漏. 为此, 我们从地震序列中随机选择了 1000 个拥有高置信度的地震 P 波初动样本. 然后, 我们请求一位地震学专家对这些样本进行了再次标注. 根据结果, 在 1000 个随机样本中, 907 个捕捉到的 P 波被判定为余震. 这个结果进一步验证了 EL-Picker 的鲁棒性. 同时, 通过对比首轮专家标注数据 (即中国地震局提供的地震 P 波标签数据), 我们发现首轮专家标记过程中, 有相当于专家标注的 125.6% 倍 P 波初动被遗漏. 这种现象启发我们进一步探索这些容易被地震专家遗忘的地震波形.

为此, 我们首先对比了首轮专家标注的地震 P 波初动与被遗漏的 P 波初动附近的振幅平均值. 图 6 第 1 行展示了两类波形在垂直方向 (Z) 的振幅变化平均值, 其中第 5 s 为 P 波初动时间点. 我们发现在首轮专家标注的地震中 (第 1 行左图), 地震波形在 P 波到后大约 15~25 s 出现剧烈波动. 相比之下, 被遗漏的地震的 P 波 (第 1 行右图), 波形相对稳定, 信噪比 (signal-to-noise ratio, SNR) 较低. 同时, 对应的 S 波初动会更加靠后, 即 25 s 后, 这表明地震中心和监测站之间的距离很长 (更多示例见图 S2). 相似的结果同样出现在我们进一步的建模分析中. 具体而言, 我们利用基于注意力机制的卷积神经网络 (recurrent attention (RA)-CNN, RA-CNN)^[49] 训练分类器判别首轮专家标注 P 波初动与被遗漏的 P 波初动. RA-CNN 通过在地震波形的频谱中对不同的区域使用不同的注意力权重, 实现比 CNN 更好的分类效果. 图 6 第 2 行展示了波形频谱中不同区域被 RA-CNN 赋予的注意力权重分布. 我们发现在 P 波初动后 15~25 s 范围内注意力权重较大, 这暗示此区域内波形应特别关注.

根据这些观察到的现象, 我们认为在人工标注过程中, 地震 S 波峰值附近波形抖动较小的地震容易被人工忽略. 尤其是在大地震发生时, 会紧随发生一系列余震, 人工标注是一项紧迫而艰巨的任务. 这时, 地震专家们倾向于优先标注出存在更大危害的较大地震, 而容易忽略信噪比较小, 需要较长时间鉴别的小地震. 同时, 我们还验证了 EL-Picker 能够有效地帮助专家们回溯容易遗漏的地震, 从而减轻地震标注任务的沉重负担.

综上所述, 我们通过多个实验分别验证了 EL-Picker 框架在地震监测应用中的鲁棒性和时效性, 灵活性和稳定性. 此外, EL-Picker 也被进一步证实能够鉴别低信噪比, 容易被人工遗漏的地震 P 波. 因此 EL-Picker 框架不仅可以减少地震实时监测中的人工负担而且可以改善监测性能.

7 总结

本文针对连续地震波形中的 P 波初动实时拾取任务, 提出了一个基于集成学习策略的技术框架, 即 EL-Picker. 具体而言, EL-Picker 由 3 个模块组成, 即触发器、分类器和精化器模块. 其中, 分类器模块借鉴集成学习策略, 实现对多个个体学习器的整合, 提升整体模型性能. 基于汶川 Ms8.0 地震的余震数据集进行的大量实验, 我们发现 EL-Picker 不仅较好地表现出 P 波初动拾取效果, 而且多诊断出 120% 被人工遗漏的地震 P 波. 同时, 我们进一步地探索了不同个体学习器对不同台站采集的地震波

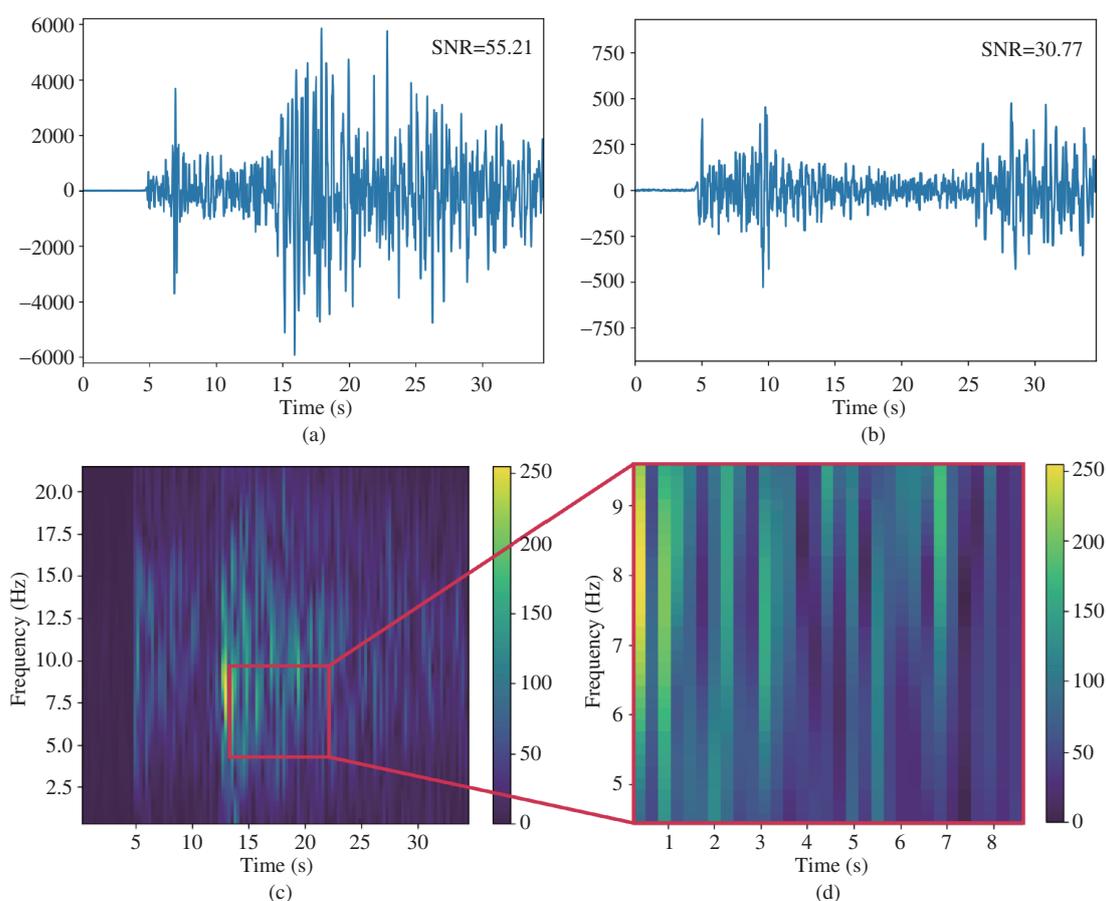


图 6 (网络版彩图) 专家标注地震 P 波与人工遗漏地震 P 波的差异性分析结果

Figure 6 (Color online) Case study on the statistics/model-based distinction between labelled and missing P-phases. (a) Expert-labelled seismic wave; (b) missing seismic wave; (c) spectral waterfall of seismic wave; (d) attention region in spectral waterfall

形的适用性以及人工遗漏的地震波形的规律特点, 这些发现清晰地验证了 EL-Picker 框架的鲁棒性、时效性、灵活性与稳定性.

致谢 感谢唐明帅对实验结果的标注, 感谢朱琛、马超、刘淇、孙莹、王宝善、王海涛等人的宝贵建议, 感谢百度人才智库其他成员的支持与鼓励.

补充材料 补充材料. 本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

- 1 Gibbons S J, Ringdal F. The detection of low magnitude seismic events using array-based waveform correlation. *Geophys J Int*, 2006, 165: 149–166
- 2 Shelly D R, Beroza G C, Ide S. Non-volcanic tremor and low-frequency earthquake swarms. *Nature*, 2007, 446: 305–307
- 3 Plenkers K, Ritter J R, Schindler M. Low signal-to-noise event detection based on waveform stacking and cross-

- correlation: application to a stimulation experiment. *J Seismol*, 2013, 17: 27–49
- 4 Saragiotis C D, Hadjileontiadis L J, Panas S M. A higher-order statistics-based phase identification of three-component seismograms in a redundant wavelet transform domain. In: *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*, 1999. 396–399
 - 5 Saragiotis C D, Hadjileontiadis L J, Panas S M. PAI-S/K: a robust automatic seismic P phase arrival identification scheme. *IEEE Trans Geosci Remote Sens*, 2002, 40: 1395–1404
 - 6 Küperkoch L, Meier T, Lee J, et al. Automated determination of P-phase arrival times at regional and local distances using higher order statistics. *Geophys J Int*, 2010, 23: 1159–1170
 - 7 Allen R. Automatic earthquake recognition and timing from single traces. *Bull Seismol Soc Amer*, 1978, 68: 1521–1532
 - 8 Allen R. Automatic phase pickers: their present use and future prospects. *Bull Seismol Soc Amer*, 1982, 72: 225–242
 - 9 Lomax A, Satriano C, Vassallo M. Automatic picker developments and optimization: filterpicker — a robust, broadband picker for real-time seismic monitoring and earthquake early warning. *Seismological Res Lett*, 2012, 83: 531–540
 - 10 Ruano A E, Madureira G, Barros O, et al. Seismic detection using support vector machines. *Neurocomputing*, 2014, 135: 273–283
 - 11 Beyreuther M, Wassermann J. Continuous earthquake detection and classification using discrete hidden Markov models. *Geophys J Int*, 2008, 175: 1055–1066
 - 12 Wang J, Teng T L. Artificial neural network-based seismic detector. *Bull Seismol Soc Amer*, 1995, 85: 308–319
 - 13 Dai H, MacBeth C. The application of back-propagation neural network to automatic picking seismic arrivals from single-component recordings. *J Geophys Res*, 1997, 102: 15105–15113
 - 14 Ross Z E, Meier M A, Hauksson E, et al. Generalized seismic phase detection with deep learning. 2018. ArXiv: 180501075
 - 15 Ross Z E, Meier M A, Hauksson E. P-wave arrival picking and first-motion polarity determination with deep learning. *J Geophys Res Solid Earth*, 2018, 123: 5120–5129
 - 16 Perol T, Gharbi M, Denolle M. Convolutional neural network for earthquake detection and location. *Sci Adv*, 2018, 4: e1700578
 - 17 Zhu W, Beroza G C. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophys J Int*, 2018, 216: 261–273
 - 18 Zhang Q, Xu T, Zhu H, et al. Aftershock detection with multi-scale description based neural network. In: *Proceedings of 2019 IEEE International Conference on Data Mining*, 2019. 886–895
 - 19 Mousavi S M, Zhu W, Sheng Y, et al. CRED: a deep residual network of convolutional and recurrent units for earthquake signal detection. *Sci Rep*, 2019, 9: 1–14
 - 20 Wiszniowski J, Plesiewicz B M, Trojanowski J. Application of real time recurrent neural network for detection of small natural earthquakes in Poland. *Acta Geophys*, 2014, 62: 469–485
 - 21 Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 785–794
 - 22 Fang L, Wu Z, Song K. SeismOlympics. *Seismological Res Lett*, 2017, 88: 1429–1430
 - 23 Brown J R, Beroza G C, Shelly D R. An autocorrelation method to detect low frequency earthquakes within tremor. *Geophys Res Lett*, 2008, 35: L16305
 - 24 Aguiar A C, Beroza G C. Pagerank for earthquakes. *Seismological Res Lett*, 2014, 85: 344–350
 - 25 Yoon C E, O'Reilly O, Bergen K J, et al. Earthquake detection through computationally efficient similarity search. *Sci Adv*, 2015, 1: e1501057
 - 26 Baer M, Kradolfer U. An automatic phase picker for local and teleseismic events. *Bull Seismol Soc Amer*, 1987, 77: 1437–1445
 - 27 Earle P S, Shearer P M. Characterization of global seismograms using an automatic-picking algorithm. *Bull Seismol Soc Amer*, 1994, 84: 366–376
 - 28 Wu Y, Wei J, Pan J, et al. Research on microseismic source locations based on deep reinforcement learning. *IEEE Access*, 2019, 7: 39962–39973
 - 29 Asim K M, Martínez-Álvarez F, Basit A, et al. Earthquake magnitude prediction in Hindukush region using machine learning techniques. *Nat Hazards*, 2017, 85: 471–486
 - 30 Riggelsen C, Ohrnberger M. A machine learning approach for improving the detection capabilities at 3C seismic

- stations. *Pure Appl Geophys*, 2014, 171: 395–411
- 31 Zhu H, Sun Y, Zhao W, et al. Rapid learning of earthquake felt area and intensity distribution with real-time search engine queries. *Sci Rep*, 2020, 10: 1–9
 - 32 Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM. In: *Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013. 273–278
 - 33 Rokach L. Ensemble-based classifiers. *Artif Intell Rev*, 2010, 33: 1–39
 - 34 Wolpert D H. Stacked generalization. *Neural Netw*, 1992, 5: 241–259
 - 35 Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000
 - 36 Breiman L. *Classification and Regression Trees*. New York: Routledge, 2017
 - 37 Ho T K. Random decision forests. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 1995. 278–282
 - 38 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*, 1997, 55: 119–139
 - 39 Cox D R. The regression analysis of binary sequences. *J R Stat Soc-Ser B (Meth)*, 1958, 20: 215–232
 - 40 Russell S J, Norvig P. *Artificial Intelligence: a Modern Approach*. New York: Pearson Education Limited, 2016
 - 41 Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of International Joint Conferences on Artificial Intelligence Organization*, Montreal, 1995. 1137–1145
 - 42 Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*, 1974, 19: 716–723
 - 43 Sleeman R, van Eck T. Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings. *Phys Earth Planet Inter*, 1999, 113: 265–275
 - 44 Leonard M, Kennett B. Multi-component autoregressive techniques for the analysis of seismograms. *Phys Earth Planet Inter*, 1999, 113: 247–263
 - 45 Leonard M. Comparison of manual and automatic onset time picking. *Bull Seismological Soc Am*, 2000, 90: 1384–1390
 - 46 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9: 1735–1780
 - 47 Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1492–1500
 - 48 Hartigan J A, Wong M A. A k-means clustering algorithm. *J Royal Stat Soc Ser C (Appl Stat)*, 1979, 28: 100–108
 - 49 Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2017. 3

El-Picker: a machine learning-enhanced robust P-phase picker for real-time seismic monitoring

Dazhong SHEN^{1†}, Qi ZHANG^{1†}, Tong XU^{1*}, Hengshu ZHU^{2*}, Wenjia ZHAO³, Zikai YIN¹, Peilun ZHOU¹, Lihua FANG⁴, Enhong CHEN^{1*} & Hui XIONG^{5*}

1. *School of Computer Science and Technology, University of Science and Technology of China, Hefei 230022, China;*

2. *Baidu Online Network Technology (Beijing) Co., Ltd., Beijing 100085, China;*

3. *Institute of Geology, China Earthquake Administration, Beijing 100029, China;*

4. *Institute of Geophysics, China Earthquake Administration, Beijing 100081, China;*

5. *Rutgers, the State University of New Jersey, Newark NJ 07102, USA*

* Corresponding author. E-mail: tongxu@ustc.edu.cn, zhuhengshu@gmail.com, cheneh@ustc.edu.cn, hxiong@rutgers.edu

† Equal contribution

Abstract Identifying the arrival times of seismic P-phases plays a significant role in real-time seismic monitoring, which provides critical guidance for emergency response activities. While considerable research has been conducted on this topic, efficiently capturing the arrival times of seismic P-phases hidden within intensively distributed and noisy seismic waves, such as those generated by the aftershocks of destructive earthquakes, remains a real challenge since most common existing methods in seismology rely on laborious expert supervision. To this end, in this paper, we present a machine learning-enhanced framework based on ensemble learning strategy, EL-Picker, for the automatic identification of seismic P-phase arrivals on continuous and massive waveforms. More specifically, EL-Picker consists of three modules, namely, Trigger, Classifier, and Refiner, and an ensemble learning strategy is exploited to integrate several machine learning classifiers. An evaluation of the aftershocks following the Ms 8.0 Wenchuan earthquake demonstrates that EL-Picker can not only achieve the best identification performance but also identify 120% more seismic P-phase arrivals as complementary data. Meanwhile, experimental results also reveal both the applicability of different machine learning models for waveforms collected from different seismic stations and the regularities of seismic P-phase arrivals that might be neglected during the manual inspection. These findings clearly validate the effectiveness, efficiency, flexibility, and stability of EL-Picker.

Keywords P-phase picker, machine learning, ensemble learning, Wenchuan aftershocks, real-time seismic monitoring



Dazhong SHEN was born in 1995. He received his B.S degree in Mathematical Sciences from University of Science and Technology of China (USTC) in 2017. He is currently working toward the Ph.D. degree in School of Computer Science and Technology, USTC. His research interests include data mining, topic model, unsupervised learning, and natural language processing.



Qi ZHANG was born in 1995. He received his B.S. degree in physical sciences from University of Science and Technology of China (USTC), Hefei, China, 2017. He is currently working toward the Ph.D. degree in School of Computer Science and Technology, USTC. His research interests include data mining, deep learning, sequential analysis, and earthquake related applications, such as earthquake detection and earthquake prediction.



Tong XU was born in 1988. He received his Ph.D. degree in University of Science and Technology of China (USTC), Hefei, China, in 2016. He is currently working as an associate professor of Anhui Province Key Laboratory of Big Data Analysis and Application, USTC.



Hengshu ZHU was born in 1986. He is currently a senior data scientist at Baidu Inc. He received his B.E. degree in 2009 and Ph.D. degree in 2014, both in computer science from University of Science and Technology of China (USTC), China. His general area of research is data mining and machine learning, with a focus on developing advanced data analysis techniques for emerging applied business research.