信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

数智创新 声至未来

DEEP IN DIALECTS, FOR FUTURE WAVE

第八届信也科技杯算法大赛

THE 8TH FINVOLUTION DATA SCIENCE COMPETITION

capybara

刘思杰 王文杰 王吉明 胡嘉豪

目录

CONTENTS

团队荣誉

- **成员介绍**

王文杰　就职于某金融科技公司，资深数据挖掘专家，毕业于上海理工大学自动化专业。

王吉明　就职于某金融科技公司，数据挖掘专家，毕业于卡迪夫大学数据科学与分析专业。

胡嘉豪　就职于某金融科技公司，数据挖掘专家，毕业于博科尼大学数据科学专业。

刘思杰　就职于某金融科技公司，数据挖掘专家，毕业于上海大学应用数学专业。

- **获得荣誉**

360数字安全公开赛-Web攻击检测与分类识别(CCF)：rank 1

电商销量预测挑战赛（科大讯飞）：rank 3

健康成人脑龄预测挑战赛（科大讯飞）：rank 4

PAKDD2021第二届阿里云智能运维算法大赛（天池）：11/1350　（< TOP1%）

企业非法集资风险预测比赛（CCF）：28/3400　（< TOP1%）

赛题理解

训练样本（10万左右）：

[{语音：$audio1$，方言（城市）：$dialect1$}，

{语音：$audio2$，方言（城市）：$dialect2$}，

……

{语音：$audion$，方言（城市）：$dialectn$}

方言间距离矩阵：

	北京	成都	郑州	武汉	广州	上海	杭州	厦门	长沙	测试方言_1	测试方言_2	…	…	测试方言_d
北京	0	32	23	34	69	68	57	79	52	X	X	X	X	X
成都	32	0	38	25	66	61	50	77	44	X	X	X	X	X
郑州	23	38				71	62	79	56	X	X			X
武汉	34	25				63	54	77	34	X	X			X
广州	69	66	69	66	0	67	68	71	68	X	X	X	X	X
上海	68	61	71	63	67	0	41	78	64	X	X	X	X	X
杭州	57	50	62	54	68	41	0	76	57	X	X	X	X	X
厦门	79	77	79	77	71	78	76	0	77	X	X	X	X	X
长沙	52	44	56	34	68	64	57	77	0	X	X	X	X	X
测试方言_1	X	X	X	X	X	X	X	X	X	0	X	X	X	X
测试方言_2	X	X	X			X	X	X	X	X				X
…	X	X	X			X	X	X	X	X	X			X
…	X	X	X	X	X	X	X	X	X	X	X	X	0	X
测试方言_d	X	X	X	X	X	X	X	X	X	X	X	X	X	0

集内-集内

集内-集外

集内-集外

集外-集外

任务目标：

给出样本对数为P，$pair_k: audio_{k,1}, audio_{k,2} \quad \forall k \in [1, P]$

预测方言pair的距离：$\hat{D}\left(audio_{k,1}, audio_{k,2}\right)$

评价函数：$\frac{1}{P}\sum_{k=1}^{P}\left|\hat{D}\left(audio_{k,1}, audio_{k,2}\right) - D\left(audio_{k,1}, audio_{k,2}\right)\right|$

复赛评测样本构成："集内-集外"，"集外-集外"

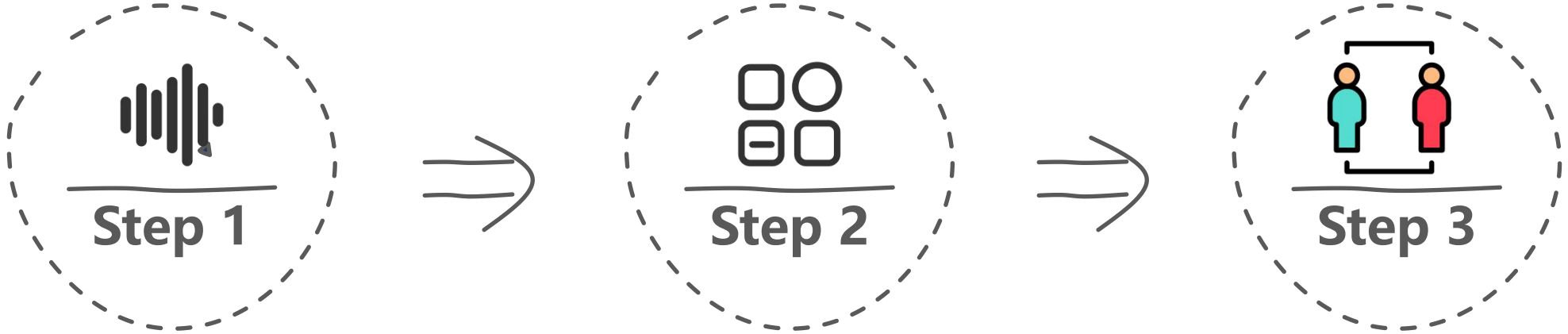
挑战：模型在集外样本上的适应能力

解题思路

考虑到任务特点，以及缺训练资源（GPU）的情况下，选择了网络提取embedding特征 + 集成树模型的模式，这种模式更适合团队协作，最大化人力、机器以及数据等资源的利用率

方案	初赛最优成绩	复赛最优成绩
神经网络回归（端到端）	9.6	14.7
神经网络回归提取emb+集成树回归	9.3	13.9
神经网络回归提取emb+集成树分类+TOP概率类别距离映射	8.9	13.2
神经网络回归提取emb+集成树分类+类别概率距离映射	7.9	11.8

最终方案



embedding提取

- ECAPA-TDNN
- Convolutional Autoencoders
- 数据增强

lightgbm分类

- ClassWeight修正
- 多折

距离映射

- 方言类别概率权重下的距离映射

特征提取

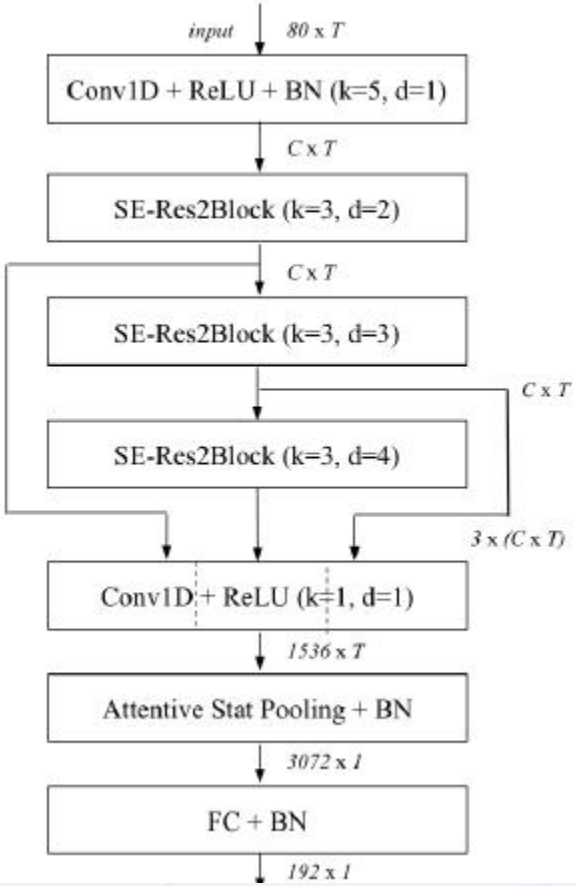
- 选择模型获取embedding：对比无监督模型、及其它一些有监督模型，这里基于效果及训练时长的考虑最终选择了ECAPA-TDNN；

- 基于性能考虑：相较传统的ECAPA-TDNN，减少了部分参数量，最终输出的embedding也减少到128维；

- 采取ECAPA-TDNN结合L1 loss训练：将训练得到的音频embedding、两两之间的L1距离拟合方言距离矩阵中对应的实际距离；

- 将获取到的embedding送入下游任务中。

ECAPA-TDNN:



数据增强

目的

- 补充更多样的训练样本，提升模型的泛化性能

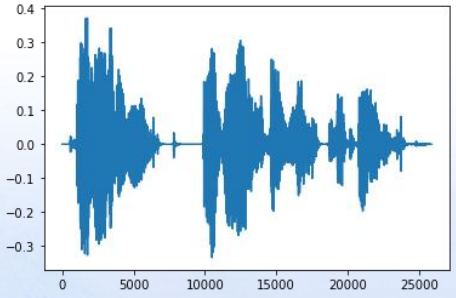
- 缓解不同方言的类别不均衡（i.e.重点针对稀缺类别方言进行音频增强）

实现

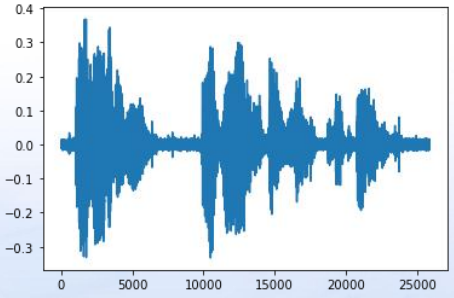
- 使用开源的音频数据增强工具 audiomentations实现数据增强
    1. 添加高斯噪音（AddGaussianNoise)
    2. 时间维度拉伸（TimeStretch)
    3. 音调拉伸（PitchShift)
    4. 时间维度滑动 (Shift)

效果图

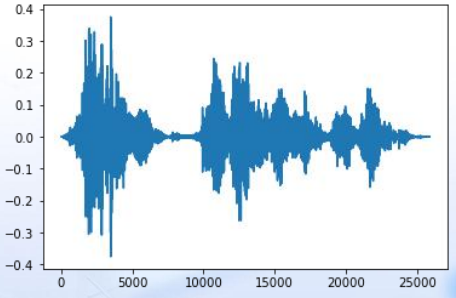
原始音频



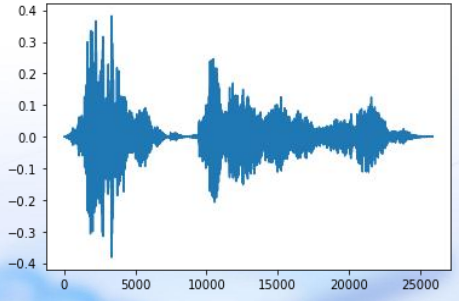
添加高斯噪声



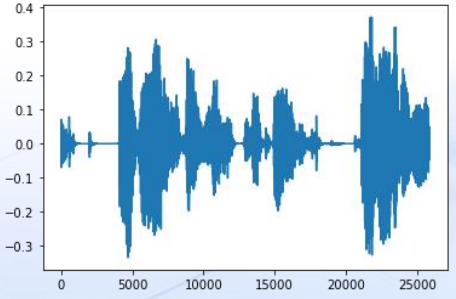
时间维度拉伸



音调拉伸



时间维度滑动



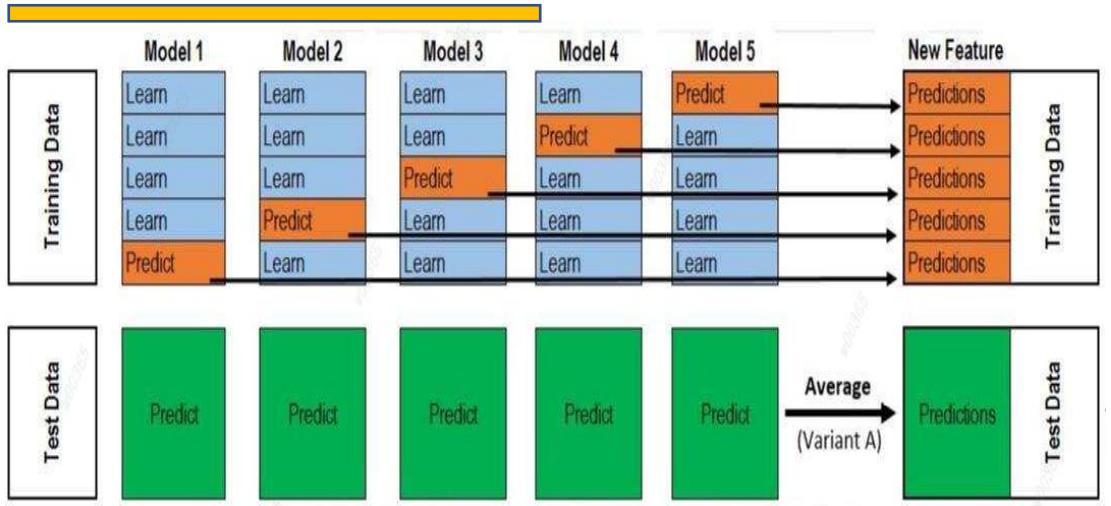
树模型分类

基于网络提取出的embedding建立多折lightgbm模型，预测方言的类别概率。

128维音频emb

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

5fold lightgbm

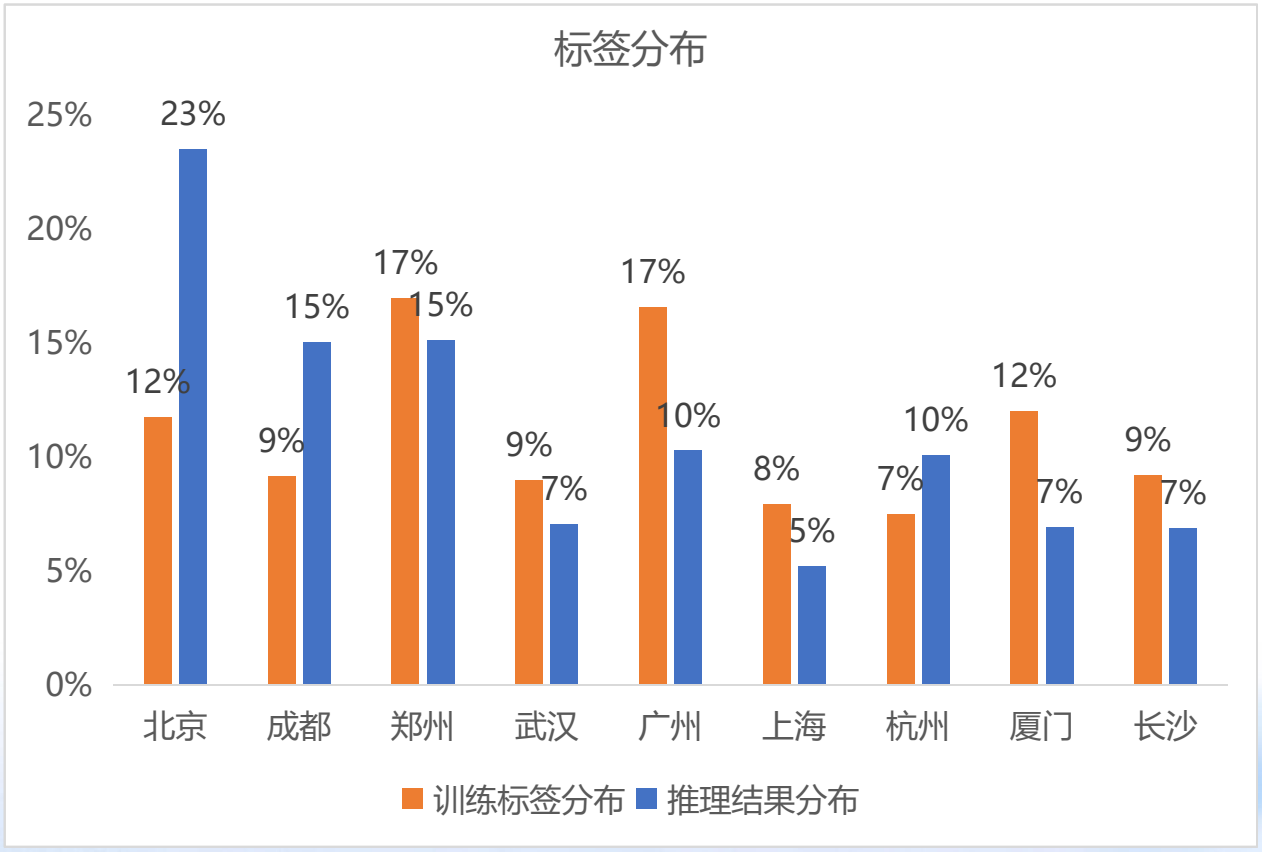


方言类别概率

pred_0	pred_1	pred_2	pred_3	pred_4	pred_5	pred_6	pred_7	pred_8
0.3%	0.6%	0.3%	0.1%	87.5%	3.3%	4.4%	1.8%	1.6%
91.9%	1.0%	7.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	99.6%	0.0%	0.0%
91.1%	0.0%	8.6%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%
1.3%	8.9%	2.1%	0.3%	0.8%	1.4%	70.7%	0.3%	14.2%
0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
0.4%	2.1%	0.1%	0.0%	0.0%	0.0%	97.3%	0.0%	0.1%
0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	99.8%	0.0%	0.0%
5.9%	1.0%	91.7%	0.0%	0.2%	0.1%	0.7%	0.2%	0.3%
0.1%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	99.5%
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
0.2%	0.1%	99.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0.1%	0.1%	99.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
99.6%	0.2%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0.0%	0.0%	0.0%	0.0%	0.0%	99.9%	0.1%	0.0%	0.0%
0.0%	0.1%	0.1%	0.0%	0.0%	0.1%	99.4%	0.0%	0.3%
0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

样本权重处理

训练集标签和测试集推理结果分布存在一定的差异性，北京和成都在测试集中明显较多，而广州、上海方言偏少，为了使训练结果像测试集靠拢，基于两者的分布差异进行了权重修正。



类别	权重系数
北京	2.00
成都	1.64
郑州	0.89
武汉	0.78
广州	0.62
上海	0.65
杭州	1.35
厦门	0.58
长沙	0.75

距离映射

由于评测样本中集外样本的存在，模型对测试样本上的识别确定性是不如训练样本，相较于使用最高概率作为方言的判定方式而言，概率权重下加工出的距离的误差明显更低（线上成绩提升0.6个点左右）。

- 给定方言a和方言b的类别概率，求解Distance(a,b)

step1. 计算方言组合概率矩阵

$$\text{Prob(a,b)} = \begin{bmatrix} p_{a1} \\ \vdots \\ p_{a9} \end{bmatrix} \otimes [p_{b1} \quad \cdots \quad p_{b9}] = \begin{bmatrix} p_{a1}p_{b1} & \cdots & p_{a1}p_{b9} \\ \vdots & \ddots & \vdots \\ p_{a9}p_{b1} & \cdots & p_{a9}p_{b9} \end{bmatrix}$$

step2. 计算最终距离

$$\text{Distance(a,b)} = \sum \text{Prob(a,b)} \odot$$

	北京	成都	郑州	武汉	广州	上海	杭州	厦门	长沙
北京	0.0	32.1	23.4	34.4	68.7	67.7	57.1	79.2	51.7
成都	32.1	0.0	38.2	24.5	65.8	60.7	49.9	76.6	43.7
郑州	23.4	38.2	0.0	40.1	68.6	70.5	61.8	79.1	55.9
武汉	34.4	24.5	40.1	0.0	66.0	62.6	54.2	77.2	34.0
广州	68.7	65.8	68.6	66.0	0.0	67.1	68.3	71.0	68.4
上海	67.7	60.7	70.5	62.6	67.1	0.0	40.7	78.0	63.9
杭州	57.1	49.9	61.8	54.2	68.3	40.7	0.0	76.2	57.3
厦门	79.2	76.6	79.1	77.2	71.0	78.0	76.2	0.0	77.0
长沙	51.7	43.7	55.9	34.0	68.4	63.9	57.3	77.0	0.0

参赛总结

缺乏训练资源的情况下，利用数据增强、权重修正、分类概率映射等方法提升了成绩

模型结构简洁且未使用任何外部数据及预训练模型

团队协作默契，发挥出了各成员特长

信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

THANK YOU