CHIP

基于Huber回归和措施分级分类编码的新冠肺炎趋势预测模型

团　队：不知叫啥队

汇报人：刘奕鑫

团队简介

刘奕鑫　华南理工大学　本科三年级

- 2020CCF-BDCI-面向数据安全治理的数据内容智能发现与分级分类（进行中...）　6st

任务说明

任务描述

在三个真实的典型区域，根据

区域防疫措施时序新闻

区域属性等特征

历史每日新冠确诊人数

预测

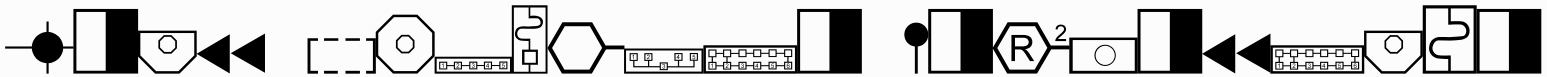
三个区域在未来X天的每日新增确诊数

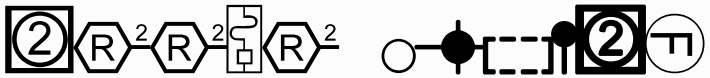
区域： A		
日期	每日确诊人数	防疫措施
2020/3/8	25	
2020/3/9	56	开始投入核酸检测
...	...	...
2020/4/31	320	建议部分区域企业可实行远程在家办公
...	...	...
2020/5/5	371	公众活动取消、禁止人群聚集、学校关闭
...	...	...
2020/5/28	600	禁止部分国家入境

总人口数	49600000
人均GDP	6000美金
65岁人口占比	8.00%
核酸测试总样本量	81146
阳性数量	5142
千人床位数	1.5

任务说明

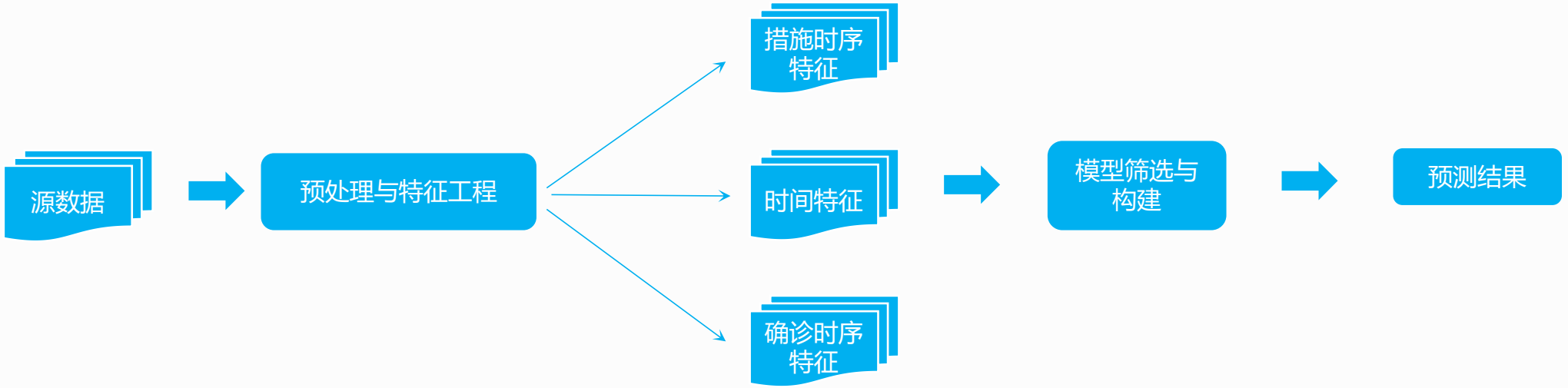
评价指标





$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right| * MAPE$$

整体方案设计



数据预处理

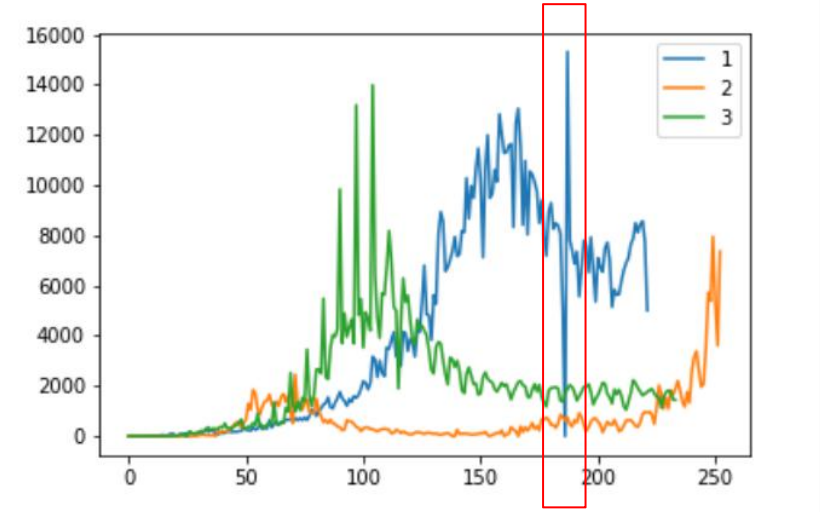
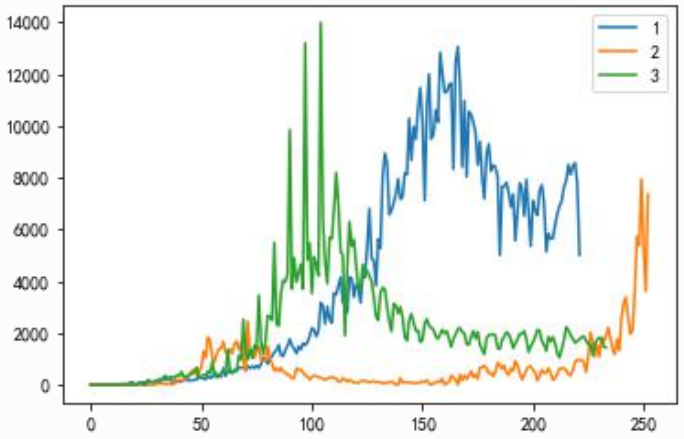


图 1. 三个区域的每日新增确诊人数曲线对比



异常数据：区域▽在○月—日确诊人数为▲人，而在○月○日为▽▽▽—人。

可能原因：原始数据统计出现了问题

处理方法：对这两天的数据进行了平均处理，两天确诊人数设置为他们的平均值

特征工程

措施特征

区域：A		
日期	每日确诊人数	防疫措施
2020/3/8	25	
2020/3/9	56	开始投入核酸检测
...	...	...
2020/4/31	320	建议部分区域企业可实行远程在家办公
...	...	...
2020/5/5	371	公众活动取消、禁止人群聚集、学校关闭
...	...	...
2020/5/28	600	禁止部分国家入境

日期	
2020-1-31	开始投入核酸检测
2020-2-24	政府建议部分区域的工作可实行远程在家办公
2020-3-12	部分地区公共活动取消、全部地区禁止100人以上的私人聚集、入境旅客需
2020-3-16	全部学校关闭
2020-3-17	全部地区公共活动取消、只允许10人以内的私人聚集
2020-3-23	禁止部分国家入境
2020-3-25	全部地区的工作场所完全关闭（除特定行业外）、全部地区公共交通限制
2020-3-26	执行部分确诊患者的密切接触人群追踪管理
2020-4-3	全部确诊患者的密切接触人群追踪管理
2020-4-14	全部地区只限制某些行业的工作人员在家办公
2020-4-24	全部地区禁止10人以下的私人聚集
2020-4-27	全部地区的工作场所完全关闭（除特定行业外）
2020-5-6	全部地区只限制某些行业的工作人员在家办公
2020-7-14	部分区域执行工作场所完全关闭（除特定行业外）、某些区域执行居家令
2020-8-6	除边境完全关闭、核酸检测以及密接人群管理外，其他防疫措施全解除。

如何对文本类型的防控措施时序数据进行特征编码？

特征工程

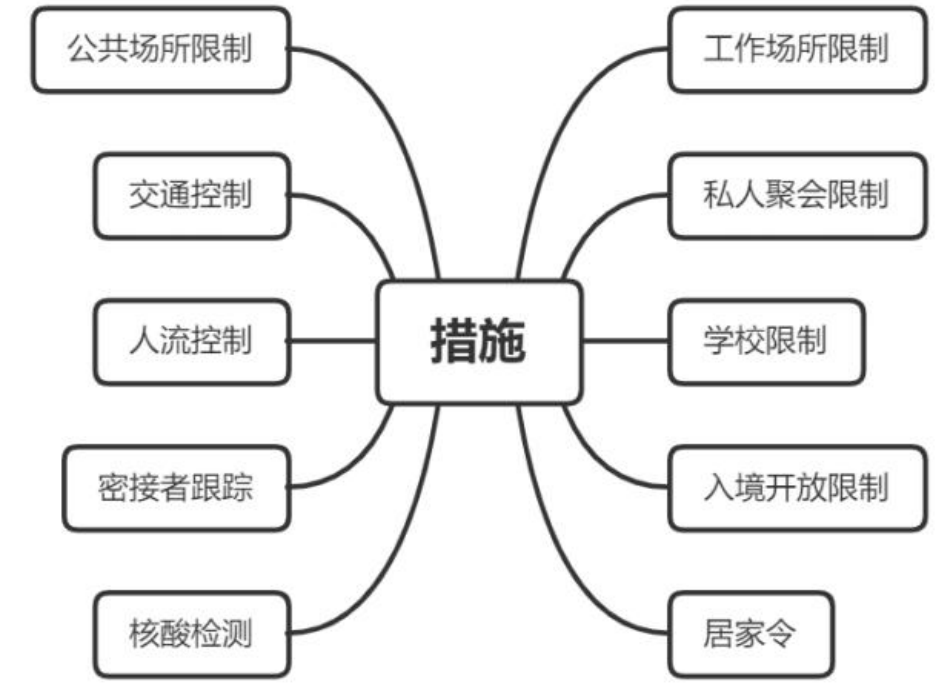


图 2. 防疫措施的分类树

防控措施编码解决方法：分级分类

▲个类别

将措施划分为工作场所限制、私人聚会限制…公共场所限制在内的▲个类别。

个级别

每个类一共有 级（无、弱、中、强）代表对应类别的措施力度，无代表该措施没有展开或者已经解除，弱到强代表该措施的力度逐渐增大。

力度归类

- 将含有“建议”字样的数据都归入弱力度的范畴
- 将含有“部分区域”、“特定”等字样的划到中等力度
- 而将含有“更严格”、“全部地区”等字样的措施划分到强力度。

[1] https://www.kaggle.com/frlemarchand/covid-19-forecasting-with-an-rnn

特征工程

其他特征

区域属性特征：我们对三个地区的数据分别进行拟合，这部分特征被我们舍弃

确诊时序特征：我们采用 [illegible] **天**作为滑动窗口长度，每一条训练数据包括该天前 [illegible] 天内的所有历史确诊人数。

日期特征：在日期特征上，我们对将**月份、日**都单独地作为了一个**类别特征**，并将时间顺序的编号也做为了一个**数值特征。**

总人口数	49600000
人均GDP	6000美金
65岁人口占比	8.00%
核酸测试总样本量	81146
阳性数量	5142
千人床位数	1.5

特征总结

训练数据

训练数据类别	数量(条)
区域1	215
区域2	246
区域3	227
总计	698

测试数据

测试数据类别	数量(条)
区域1	7
区域2	7
区域3	7
总计	21

留出最后×天作为本地测试数据

主要特征

类别	特征	说明
时间特征(x3)	day_count	日期的编号
	month	月份
	day	天
措施特征(x10)	nat_level	当前核酸检测措施等级
	school_level	当前学校限制等级
	…	…
	traffic_control_level	当前交通控制等级
历史确诊时序特征(x14)	Lag_1	前1天确诊人数
	Lag_2	前2天确诊人数
	…	…
	Lag_14	前14天确诊人数

设样本的数目为 ⅞ 则我们的输入维度为 ○/ ◑X㊦

模型筛选

在三个区域的训练数据上进行了10折交叉验证，筛选出了每个区域MAPE指标下TOP5的模型

区域1

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
huber	Huber Regressor	567.0685	849652.4673	889.0579	0.9352	0.3205	0.2331	0.0090
rf	Random Forest Regressor	526.5350	826136.1429	864.7473	0.9408	0.3329	0.2416	0.0430
et	Extra Trees Regressor	532.5723	760042.3718	822.5109	0.9468	0.3268	0.2530	0.0370
knn	K Neighbors Regressor	557.1974	882906.6844	903.5663	0.9367	0.3678	0.2545	0.0080
xgboost	Extreme Gradient Boosting	567.1734	814589.9250	877.5864	0.9394	0.3385	0.2667	0.1810

区域2

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
huber	Huber Regressor	153.3308	75197.3273	259.6272	0.7625	0.8429	0.3297	0.0120
et	Extra Trees Regressor	147.4039	56904.6101	227.6383	0.8189	0.8397	0.3487	0.0440
catboost	CatBoost Regressor	151.7660	65453.4926	240.0064	0.8090	0.9933	0.3647	1.6480
rf	Random Forest Regressor	163.9137	76644.9816	257.9366	0.7696	0.9220	0.3684	0.0460
knn	K Neighbors Regressor	187.9483	100438.6549	297.4251	0.7103	0.8835	0.3733	0.0090

区域3

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
huber	Huber Regressor	340.5845	484996.1944	617.9748	0.8556	0.4566	0.1963	0.0110
knn	K Neighbors Regressor	372.4855	766918.7922	755.3770	0.7978	0.5592	0.2143	0.0090
rf	Random Forest Regressor	375.5096	588092.7919	690.7730	0.8195	0.5016	0.2170	0.0550
et	Extra Trees Regressor	344.4651	504059.8308	636.5573	0.8311	0.5411	0.2286	0.0400
xgboost	Extreme Gradient Boosting	425.0086	646378.0383	740.0350	0.8070	0.5722	0.2688	0.0990

Huber回归

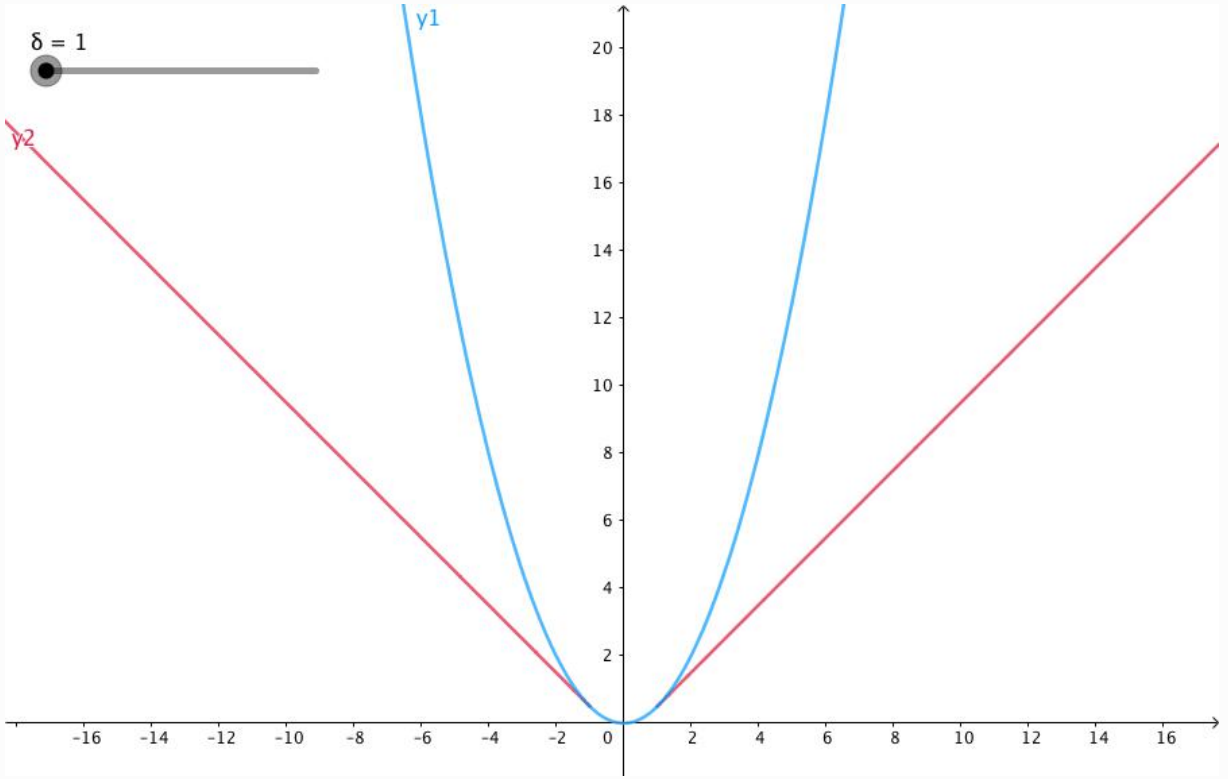
Huber回归（最小二乘回归的一种变体）

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2, & \text{for } |a| \le \delta \\ \delta \cdot \left(|a| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}$$

其中， $a = \hat{y} - y$， δ为可学习参数。

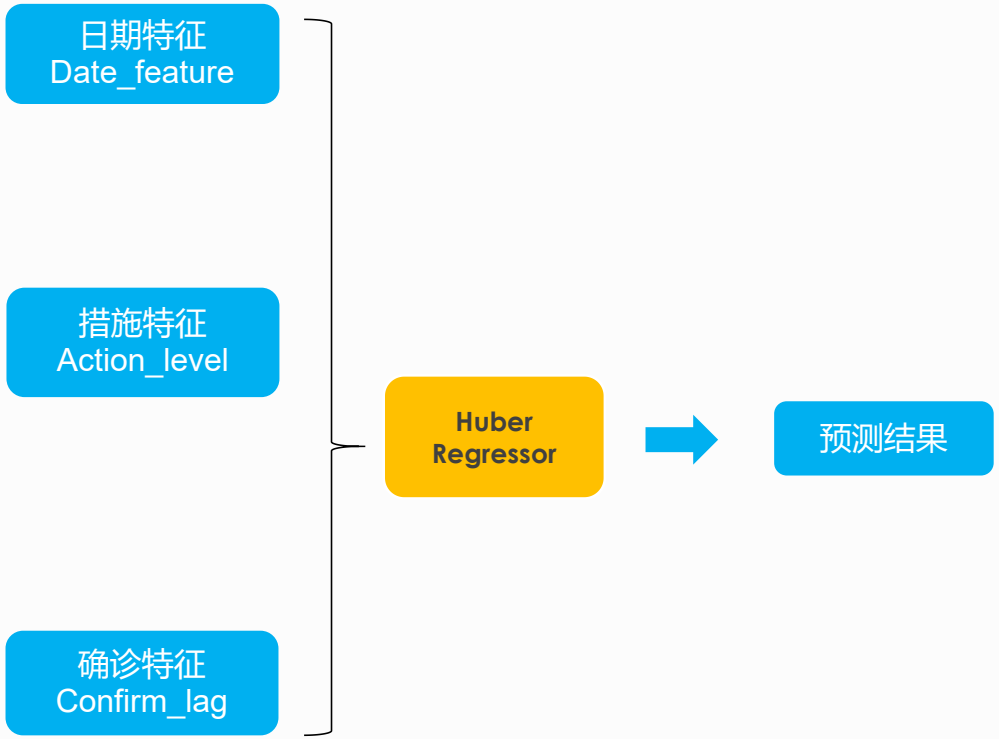
解释：
当预测偏差小于 δ 时，它采用平方误差
当预测偏差大于 δ 时，采用的线性误差。

优点：
能够比较好地降低模型对异常值的敏感度



最终模型

在测试集上，模型集成和融合效果都没有精调的单Huber模型表现地好



模型效果

日期	新增确诊（区域 1）	新增确诊（区域 2）	新增确诊（区域 3）
2020-10-14	5014	7359	1448
2020-10-15	6675	6177	1481
2020-10-16	7164	5280	1498
2020-10-17	7457	6220	1874
2020-10-18	7509	6277	1706
2020-10-19	7325	6134	1675
2020-10-20	7298	6134	1573

地域	本地测试集MAPE
区域1	0.1729
区域2	0.2423
区域3	0.1039
总计(平均)	**0.1730**

线上测试集得分（换算区域平均MAPE）：0.2061

还有什么能做?

1

针对措施文本时序数据的自动特征提取模型

2

补充更多数据

3

更细致的特征工程

THANKS

团　队：不知叫啥队

汇报人：刘奕鑫