

重庆市 CDC 疾病预测项目

——手足口病发病人数的预测报告

作者：申万祥

时间：2017 年 6 月-2017 年 7 月

目录

- 一、背景及介绍.....3
- 二、数据来源与分析.....3
 - 2.1 手足口病数据.....3
 - 2.2 天气及天气预报数据.....3
 - 2.3 舆情数据.....4
- 三、预测方案及工具.....4
 - 3.1 模型的架构.....4
 - 1) 单模型构架思路.....4
 - 2) 组合模型构架思路.....5
 - 3.2 所用工具.....5
- 四、特征抽取与分析.....6
 - 4.1 特征抽取、处理与分析.....6
 - 4.2 特征选择.....6
- 五、模型评价指标.....7
 - 5.1 模型预测的精确度评分.....7
 - 5.2 模型的滞后性打分.....7
- 六、建立模型.....8
 - 6.1 模型总体流程.....8
 - 6.2 单模型预测.....8
 - 6.3 组合网络模型预测.....9
 - 1) ： 基于偏差的组合模型.....9
 - 2) ： 输入输出组合模型.....9
- 七、结果与分析.....10
 - 7.1 发病人数序列分析.....10
 - 7.2 特征相关性分析.....10
 - 7.3 模型建模过程.....12
 - 7.4 模型结果与讨论.....13
 - 1. 单模型结果.....13
 - 2. 组合模型结果.....14
- 八、结论.....18
- 九、参考文献.....18
- 十、代码模块化与说明.....18

一、背景及介绍

手足口病(Hand foot mouth disease, HFMD) 是由肠道病毒引起的常见传染病，在临床上以手、足、口腔疱疹为主要特征，故通称为手足口病，多发生于 5 岁以下的婴幼儿，所以常被称作"小儿手足口病"。HFMD 全年均可有发病，但 3~11 月份多见，6~8 月份为高峰期。传播速度极快，传播范围极广，具有周期流行的规律，一般 2~3 年流行一次。根据百度百科和维基百科，HFMD 的发病潜伏期约 1 周（多为 2~10 天，平均 3~5 天）。

二、数据来源与分析

2.1 手足口病数据

重庆市手足口病数据来自重庆市 CDC 部门，涉及到发病人数和就诊人数。其中就诊人数为医疗机构上报的就诊人数。涉及的发病地区主要有重庆，也有其他地区如北京，陕西，江西等，但是家庭住址全部为重庆地区的各个区县。原始的数据为 2012 年到 2016 年，分别为 5 个 excel 保存。因为原始的只有地区

编码和住址编码，所以通过编码与地址对码，以及数据整合，最终生成一个 csv

文件，包含所有年份的数据，以及补充的报告区县，发病区县，家庭住区县

(文件名为: **szk_data.csv**)

2.2 天气及天气预报数据

重庆市的天气及天气预报数据通过爬虫获取，文件名为（**重庆实际天气.csv, 重庆天气预报.csv**）。天气数据分为白天和夜间的数据，在后面的特征转换盒抽取中，将天气描述性文字量化，将白天和夜间平均。

2.3 舆情数据

目前的舆情数据主要是百度指数（文件名：**baidu_zhishumobile.txt, baidu_zhishupc.txt**），由于该指数缺失值、奇异值过多，所以本次建模预测的时候暂时没有采用这部分数据，但是数据经过整合、填充，目前最新的数据文件名为：。

三、预测方案及工具

3.1 模型的架构

所有的模型采用滚动预测的方法，进行一步预测（前一周预测后一周的）。

所以随着滚动，训练集数据会逐渐增加。主要采用单模型和组合模型进行滚动预测。

1) 单模型构架思路

单模型包括自动滞后一周模型（EWA 模型），这是参照基准模型。最终的模型的效果应该远远比基准的效果要好，否则没有多大意义。其次是季节性 ARIMA 模型和非季节性 ARIMA 模型，这个模型利用时序数据，加入或者不加天气因子，以及天气预报因子。再次是有监督的 SVR 模型，GBDT 模型以及 DT-Adaboost 模型，该模型加入的特征包括：发病人数本身的滞后 x 周， n 阶差分滞后 x 周，气象特征滞后 x 周，时间特征不做任何滞后，可以看作是时效特征。所有单模型构建过程如图 1。

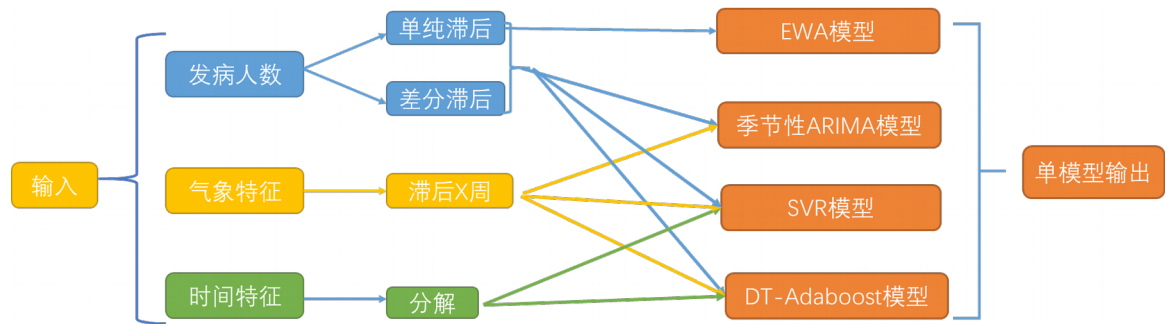


图 1. 单模型构建思路流程

2) 组合模型构架思路

组合模型思路大致分为两类，第一类是最终的输出=模型 1 的结果+模型 2 的结果，第二类是最终的输出=模型 2 的输出（其中模型 1 的输出是模型 2 的输入）。因此，第一类组合模型可以称为**基于偏差的组合模型**，这种方法主要是：模型 2 预测模型 1 的偏差，最后的结果就是模型 1 的结果加上模型 2 的结果。第二类组合模型可以称为**输出输入组合模型**，这种方法主要是：模型 2 的特征来源包括了模型 1 的输出。两类组合模型的特征输入如下图 2 所示

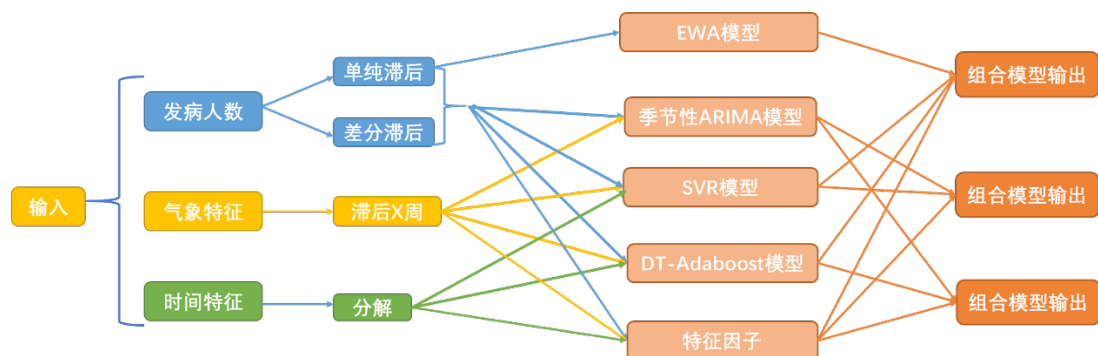


图 2. 组合模型构建思路流程

3.2 所用工具

所有数据分析，处理，特征选择，构建模型，可视化和打分函数都用 python 语言，经过后续封装，可以直接调用 **swx_ts_prediction** 这个包使用，运行环境为 python 2.7，其中包括 **feature processing**, **evaluation**, **model**, **visualization** 这四个小模块。该包依赖的主要的库有：
pandas, numpy, scipy, statsmodels, sklearn, matplotlib 和 seaborn。调用该包时，确保加入环境变量，可以通过 `sys.path.append(r"D:\...")` 实现，详细信息请阅读 `readme.rst` 文件。

四、特征抽取与分析

4.1 特征抽取、处理与分析

1) 真实天气及天气预报数据包括 Categorical 和 Numerical 数据，其中 Categorical 数据分解为雨天，晴天，云天，雷天，雾天，阴天，雪天，然后根

据各种天气有描述性程度，附上不同的权重。Numerical 数据主要是整合，差分
和滞后；

2) 手足口病发病人数数据，主要采用差分，滞后进行特征抽取。差分滞后
由函数，其中对应的时间分解为年、月、日、周、季这 5 个特征；

3) 舆情数据主要是 Numerical 数据，但是有缺失值，主要是采用前后一天
数值进行缺失补充。

抽取到以上特征之后，进行差分滞后，单纯滞后，对特征做分布分析和相
关性分析，采用 skew 方法将 skewness 大于 0.75 的取以对数，使得数据更加近
似成为正态分布。对于高度线性相关(相关系数大于 0.95)的特征保留其中的一个

4.2 特征选择

1) 去除斜方差小于 0.8 的特征；

2) 单变量选择法：利用皮尔逊相关系数筛选线性相关性较大的特征（与训
练集合数据的发病人数相关性大于 0.2），此外，一些完全线性不相关的但可能
是导致疾病的因素的特征也被添加进来；

3) 在线性 SVR 模型中采用循环递归消除法选择特征（RFE）：具体做法是：
反复的构建模型（SVR 回归模型）然后交叉验证选出最好的特征（根据系数来
选），把选出来的特征放到一遍，然后在剩余的特征上重复这个过程，直到所

有特征都遍历了。

五、模型评价指标

5.1 模型预测的精确度评分

- 1) . 平均绝对百分偏差率 (MAPE) : $MAPE = (\sum(|X-Y|)/X) * 100\% / N$
- 2) . 平均绝对偏差 (MAE) : $MAE = \sum|X-Y| / N$
- 3) 赤池信息量准则 (AIC) 和贝叶斯信息准则 (BIC) : $AIC = 2 * \log-Likelihood + 2 * K$; $BIC = 2 * \log-Likelihood + K * \log(N)$ (K 为参数个数)

5.2 模型的滞后性打分

此外，为了衡量模型在未来预测的上升或者下降走势的准确率，以及在疾病人数上涨的场景的预警率，根据上升和下降，将变化转换为分类问题的评价指标，包括以下 4 个指标（越高越好）：

- 1) 波峰，波谷皮尔逊相关系数 (PCCP, PCCV) : 波峰活波谷处皮尔逊相关系数的平均值；
- 2) 一步预测涨跌准确率 (ACC) : $ACC = (TN + TP) / (TN + TP + FN + FP)$

3) 上升趋势提前预警率 (SEN) : $SEN = TP / (TP + FN)$

4) 下降趋势提前预警率 (SPE) : $SPE = TN / (TN + FP)$

六、建立模型

6.1 模型总体流程

1) 使用滑动窗口+回归模型的方式完成建模, 对过去 K 周数据训练建立回

归模型, 预测下一周趋势;

2) 训练窗口随时间持续滑动(训练集逐渐增加), 以保证模型的时效性;

3) 每次增加一定步长的训练集样本, 就对模型的最佳参数进行格点搜索,

更新模型的最佳参数 (对于 ARIMA 模型, 采用 BIC, AIC 作为参数寻

优的指标, 对于其他监督性学习模型, 5-折交叉验证训练集, 然后采用

MAE 或者 R^2 作为寻优标准)。

6.2 单模型预测

1) : EWA 平移一周: 将发病人数平移一周;

2) : SARIMA 单模型: 不加任何外部因子, 直接采用病例数作为输入、输

出，优化的参数有 p, d, q ；

3) : SARIMA 单模型加外部因子：分别添加滞后的真实和预报的天气因子，优化的参数有 p, d, q 和 P, D, Q, s ；

4) : SVR 单模型加外部因子：除了滞后、差分的天气因子，还添加时间因子，病例数滞后因子，病例数差分滞后因子，SVR 中分别采用 RBF 非线性核函数和线性核函数，优化的参数有：C, epsilon, gamma (RBF 核)，loss (线性核)；

5) : GBDT 单模型加外部因子：GBDT 外部因子和 SVR 相同,GBDT 模型每一步主要调节 loss, learning_rate, n_estimators, max_depth 这四个参数；

6) : Adaboost 模型加外部因子：外部因子和 SVR 相同，模型使用 DecisionTreeRegressor 作为单个叠加器，优化的参数有：learning_rate, n_estimators, max_depth, loss。

6.3 组合网络模型预测

1) : 基于偏差的组合模型

EWA 偏差组合模型：预测 EWA 与真实值的差异，然后加上 EWA 的预测值，为最终的输出，组合模型的类型有：

EWA+SVR, EWA+Adaboost, EWA+GBDT;

SARIMA 偏差组合模型：预测 SARIMA 与真实值的差异，然后加上 SARIMA 的预测值，根据 EWA 偏差组合模型结果，只做了 SARIMA+SVR 的组合模型；

在以上偏差预测过程中，偏差项需要与因子做进一步的相关性分析，选取相关性较大的因子作为偏差项的输入。

2)：输入输出组合模型

输出输入组合模型是：第一步：整合天气因子等作为一个相关性较强的特征，第二步，该特征结合单模型输出的结果，以及天气因子等一起作为输入得到一个最终的结果。初步采用的组合模型的方式为：

SARIMA + Adaboost + 天气因子——>SVR

主要是考虑到如果使用相同的组合模型，会造成误差的叠加，而使用不同的组合模型则会取长补短，最终用 SVR 作为最后模型是因为，SVR 与其

他方法相比，采用 RFE 特征选择算法，可以选则性的加入特征，从而提升预测精度。

七、结果与分析

7.1 发病人数序列分析

发病人数的数据分布如图 6 所示，数据是重庆发病人数整体的统计结果。发病人数呈现周期性变化，趋势分解可以看出周期约为 180 天，约为 25~26 周。

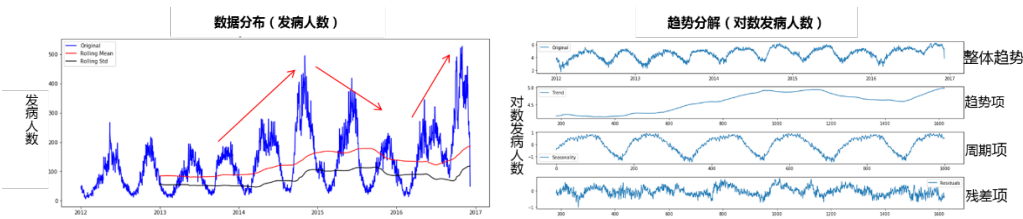


图 6. 序列分布

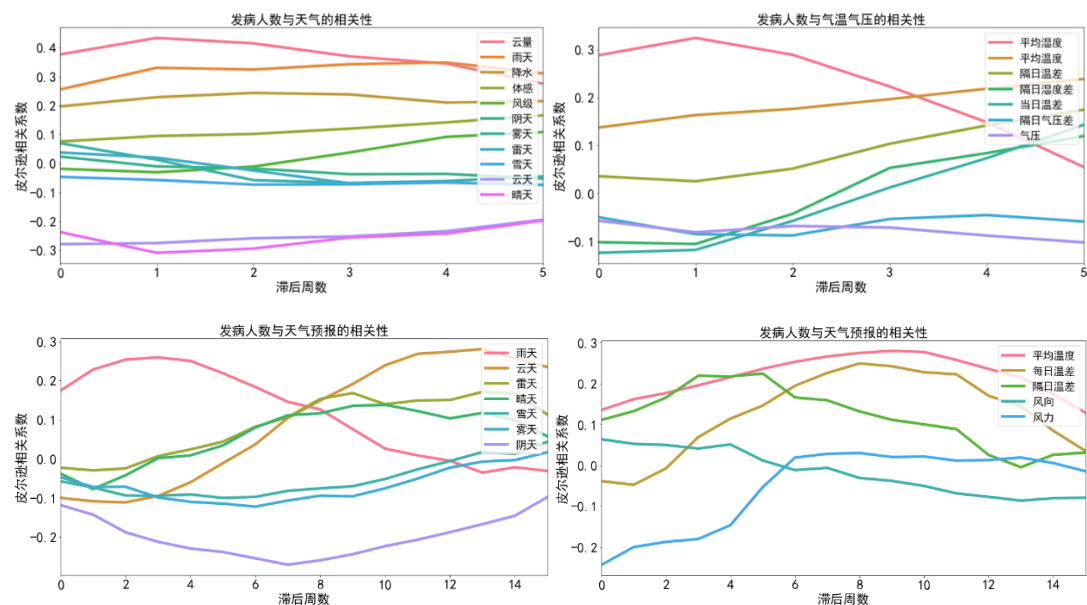
图 7. 序列趋势分解

7.2 特征相关性分析

1) 真实的天气特征滞后相关性见图 5，真实天气特征滞后一周雨发病人数有明

显的相关性上升趋势，说明前一周的天气对发病人数影响较大，其中，前一周的云量、雨天与发病人数呈显著正相关，而前一周的晴天和云天与发病人数呈

显著负相关，对应前一周的平均湿度呈显著负相关 ($p < 0.001$)。与真实天气



以上结果说明天气特征对手足口病爆发的导火线性因素，而手足口病爆发

2) 发病人数序列自相关性和偏差相关性见图 6 和图 7: 自相关和偏差相关性都

24~26 周前的数据，差分后最强的相关性为：差分 11 周，滞后一周的数据，相关系数都在 0.8 以上 ($p \ll 0.001$)。

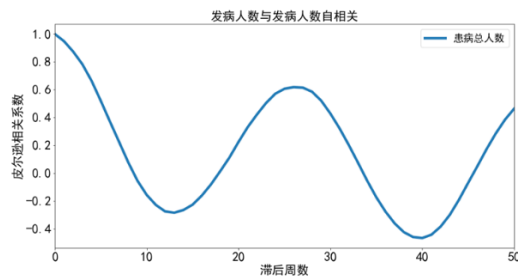


图 6. 序列不同滞后自相关

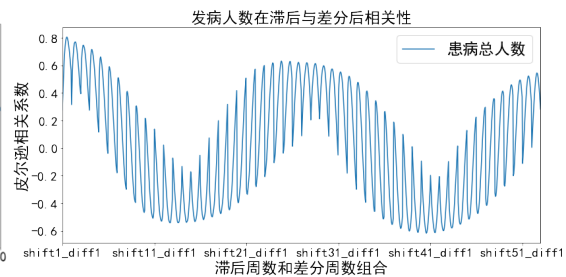


图 7. 序列不同差分后自相关

3) 舆情因子的相关性分析

未完成

7.3 模型建模过程

1) SARIMA 单模型：SARIMA 模型的建立，在模型选择的标准方面，采用训练集中最小的 BIC 和 AIC 值来选择模型，即确定 p, d, q 和 P, D, Q ，其中因为手足口病小周期接近 12~13 周，所以先尝试使用 s 为 13，发现取 12 效果较好，所以后续优化就固定了 s 为 12，然后优化 p, d, q 。因为加入不同的训练集，最优的模型可能不一样，所以设置了步长为 30（因为 SARIMA 速度较慢，所以步长没有选择为 1），选择一次模型。图 8 显示了建模过程中，选择不同模型的时候 AIC

和 BIC 的变化，可以看得出，随着训练集的增加，最小的 BIC 和 AIC 也逐渐变小。

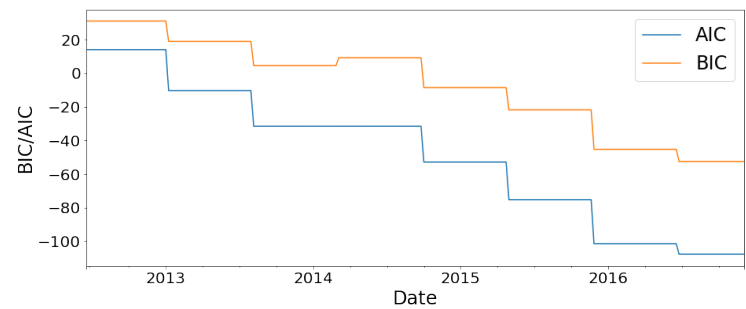


图 8. 滚动预测最小的 AIC 和 BIC 的变化

2) 其他监督学习单模型

其他监督性学习模型在参数优化的过程中，打分函数可以选择 R 平方或者负的 MAE，原始代码默认为负的 MAE。在随着滚动窗口的前进，训练集的增加，5 折交叉验证的 R2 都会越来越高，如图 9 所示（预测偏差组合模型），MAE 越来越低，其变化如图 10 所示（输入输出组合模型）。

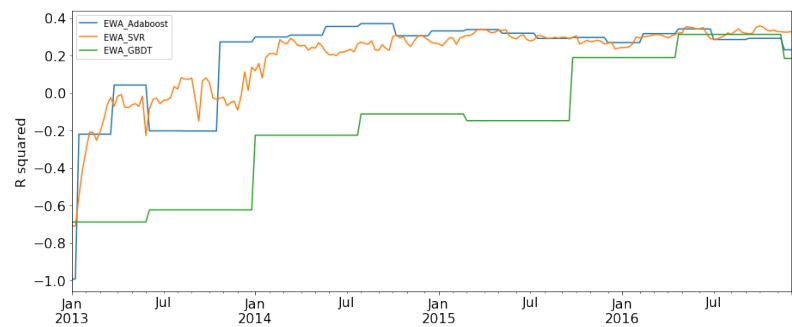


图 8. 偏差组合模型中单模型的滚动预测 R2 的变化

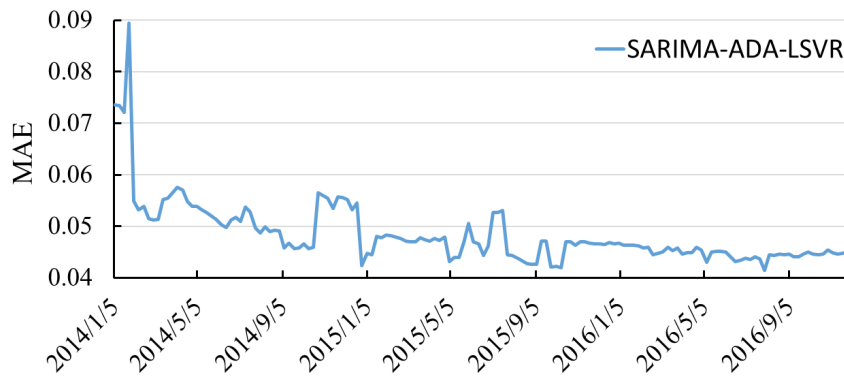


图 10. 输入输出组合模型滚动预测 MAE 的变化

7.4 模型结果与讨论

1. 单模型结果

1) 模型的最终对比结果使用的是 2014 到 2016 年的预测结果，见表 1。其中，效果最好的单模型是 SARIMA+ 部分天气因子单模型，MAPE 为 0.144，ACC 可以达到 0.73，PCCP 可以到 0.79。相比较加入全部天气因子，加入部分因子使得结果改善（0.144 vs. 0.156），去掉的因子就是协方差小于 0.8 的特征，这说明在 SARIMA 中，使用协方差过滤因子是非常有效的。

2) 其他单模型中，RBF-SVR 效果最好，在 RBF-SVR 中采用了特征选择，每次滑动的时候选择 PCC 大于 0.0001 的特征。选取相关性较小的特征是因为同

时考虑非线性相关特征。在其他单模型中，过滤协方差较小的特征提升不明显甚至变差。这说明在其他监督性学习模型中（ADA, GBDT, RFE-LSVR），没有必要首先过滤特征，因为这些方法自带有特征选择；在 SVR 中，使用 RBF 核函数性能比线性核函数性能好很多（0.177 vs. 0.183）。值得注意的是，这些单模型的 MAPE 效果虽然没 SARIMA 的好，但是在滞后性指标上（表 1），这些模型的效果整体较好，说明这些单模型有助于改善模型的滞后性（可参见图 11，图 12）。

以上结果和文献报道较为一致[1]，RBF-SVR 的性能也比较好。同时文献报道了 SARIMA 结合卡尔曼滤波（Kalman filter）效果最好。其他文献[2]中也有相似的结果，说明卡尔曼滤波的确能改善 SARIMA 模型，后续可以探索和尝试这种方法。

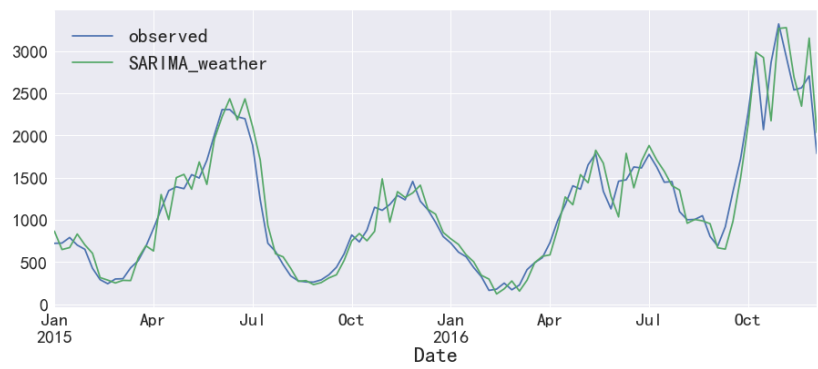


图 11. SARIMA+weather 单模型的预测效果

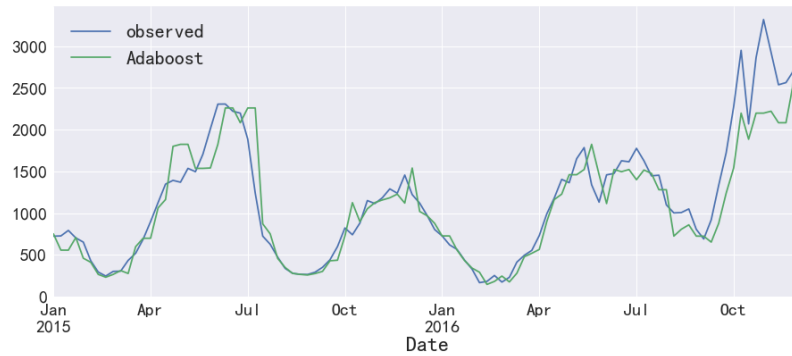


图 12. Adaboost 单模型的预测

2. 组合模型结果

在基于偏差的组合模型中，**初始模型（EWA 或者 SARIMA）总是加上 SVR 的效果好**，并且 SVR 相比于其他模型，速度最快。目前该类组合模型最好的是 SARIMA+SVR 组合，和单模型 SARIMA+部分天气因子一样都是 0.144。SARIMA+SVR 组合是基于 SARIMA 不加任何天气因子做得结果（0.156），后续可以基于 SARIMA+部分天气因子，然后加上 SVR 提高预测精度，其中 SVR 可以选择舆情因子等；

在基于输出输入组合模型中，SARIMA + ADA +SVR 是所有模型效果中最好的模型。MAPE 可以达到 0.132（三年的预测效果见图 13），同时滞后性也得到了很大的改善。分析原因是：使用了 SARIMA 的输出，确保了结果不会太差，使用 Adaboost 整合的天气因子，改善滞后性，同时加入天气因子，再利用 RFE 特征选择算法每次有选择利用不同的特征。这种三组合模型改善的主要原因是集成了以上单模型的各自优点。

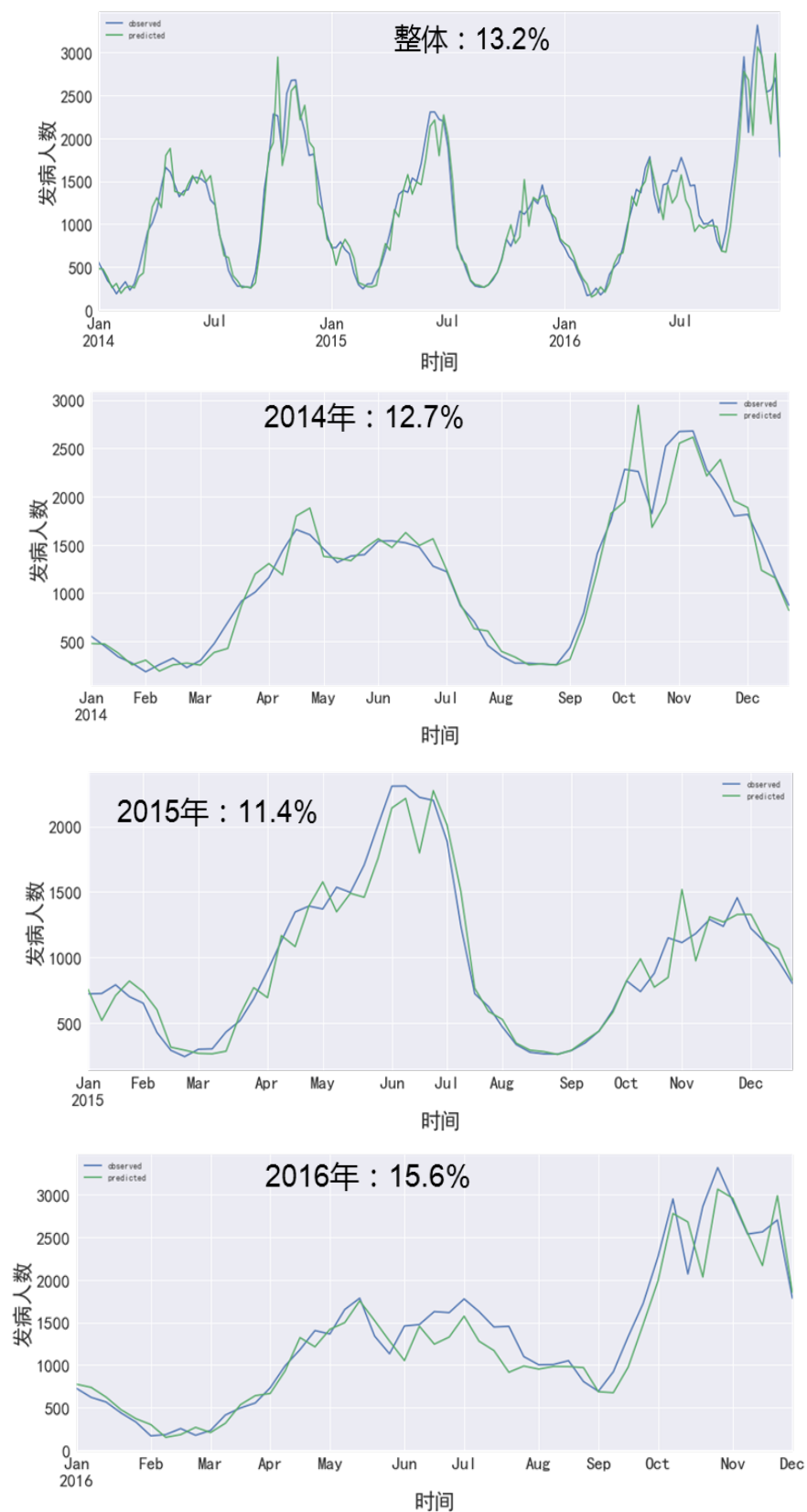


图 13 SARIMA + ADA +LSVR 整体金额分年的预测效果

表 1 单模型和组合模型预测结果对比（预测的时间段为：2014~2016 年）

	Model	MAP E	ACC	SEN	PCCP	PCCV	TIME(s)
单 模 型	EWA	0.188	0.684	0.680	0.527	0.954	--
	SARIMA	0.158	0.68	0.69	0.56	0.96	130
	SARIMA+ forecast	0.165	0.645	0.667	0.54	0.954	145
	SARIMA+ weather	0.144	0.730	0.79	0.79	0.954	145
	SVR(Linear,R FE)	0.183	0.697	0.68	0.48	0.96	30
	SVR(RBF)	0.177	0.71	0.70	0.71	0.96	40
	GBDT	0.211	0.736	0.76	0.63	0.945	120
	ADA	0.202	0.72	0.786	0.53	0.93	180
组 合 模 型	EWA + GBDT	0.159	0.64	0.64	0.48	0.92	
	EWA + ADA	0.156	0.65	0.65	0.53	0.96	
	EWA + SVR	0.145	0.66	0.65	0.55	0.97	
	SARIMA + SVR	0.144	0.77	0.77	0.64	0.96	
	SARIMA + ADA +LSVR	0.132	0.71	0.75	0.79	0.954	

【注】：组合模型中，浅色的是基于偏差的组合，深色的是基于输出、输入的

组合

八、结论

通过使用单模型，组合模型的对比，发现：












- 1) 在单模型中，他们各有利弊。SARIMA 的预测结果的偏差总是较小(最好的是 0.144)，但是滞后性比较严重，其他模型的预测偏差较大，但是滞后性相比于 SARIMA 有所改善；
- 2) 采用组合模型集成他们各自优缺点，发现 SARIMA+ADA+LSVR 效果最好，MAPE 为 0.132，此外，SARIMA+SVR 效果也不错，MAPE 为 0.144。以上结果说明组合模型有利于模型精度的提升，同时为了避免模型之间误差的叠加，应该采用不同的模型取长补短来组合。

九、参考文献

- [1]. Lippi M, Bertini M, Frasconi P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(2): 871-882.
- [2]. Liu H, Tian H, Li Y. Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction[J]. Applied Energy, 2012, 98: 415-424.

十、代码模块化与说明

代码主要分为 4 个部分：

PY 文件 (5)			
 _init_.py	2016/12/30 2:50	PY 文件	1 KB
 evaluation.py	2017/6/29 10:56	PY 文件	7 KB
 feature_preprocessing.py	2017/6/29 10:14	PY 文件	4 KB
 model.py	2017/6/29 10:45	PY 文件	18 KB
 visualization.py	2017/6/29 11:05	PY 文件	15 KB
PYC 文件 (5)			
 _init_.pyc	2017/6/15 17:46	PYC 文件	1 KB
 evaluation.pyc	2017/6/29 11:03	PYC 文件	7 KB
 feature_preprocessing.pyc	2017/6/29 10:15	PYC 文件	4 KB
 model.pyc	2017/6/29 11:03	PYC 文件	14 KB
 visualization.pyc	2017/6/29 11:06	PYC 文件	15 KB
RST 文件 (1)			
 readme.rst	2017/6/29 11:25	RST 文件	1 KB

1. feature_preprocessing 模块：

- remove_low_variance_features(data_frame,var = 0.8)函数：去除方差较小特征，VarianceThreshold 为 0.8
- check_skew_log(df,alpha = 0.75)函数：将所有特征近似为正态分布，skewed 阈值为 0.75
- diff_shift_lag(df,diff_lag = 10,shift_lag = 2)函数：将时序滞后，差分，得到新的特征，diff_lag 为差分最大阶数，shift_lag 为滞后最大阶数
- series_to_supervised(data, n_in=1, n_out=1, dropnan=True)函数：将时序滞后 n_in 次得到新的特征

2. model 模块：

- grid_search_ARMA_para(ts, pdqrange = (0,2))函数：用于 SARIMA 模型不加外部因子情形下的格点搜索，同时也可以搜索 s，按需修改
- ARMA_train_predict_rolling(ts_log, roll_begin_index = 5, para_search_step = 20)函数：SARIMA 不加外部因子模型，roll_begin_index 表示起始滑动的索引，para_search_step 为格点搜索的步长；

- c) `grid_search_ARMA_Ex_para(ts,ex_train, pdqrange = (0,2))`函数，与 `grid_search_ARMA_para` 相同，只不过是加了外部因子（有时间需要加工整合一下）
- d) `ARMA_Ex_train_predict_rolling(ts_log, ex_vectors, roll_begin_index = 51, para_search_step = 20, pdqrange = (0,2))`函数：加外部因子的 SARIMA 模型，`ex_vectors` 为外部因子
- e) `my_score_(ground_truth, predictions)`函数：自定义打分函数，可以定义好加入模型中，在参数寻优时采用此打分函数，也可以采用默认的如 `r2`，`MAE` 等；
- f) `Adaboost_train_predict_rolling(dff, roll_begin_index = 51, para_search_step = 20, cv_score = 'neg_mean_absolute_error', cv = 5)`函数：ADA 模型的函数，`cv_score` 即为格点搜索打分函数，优化的参数请在源码中按需修改，这部分由于时间原因没有加进来。CV 是交叉验证次数；
- g) `GBDT_train_predict_rolling(dff, roll_begin_index = 51, para_search_step = 20, cv_score = 'neg_mean_absolute_error', cv = 5)`函数，与上面的函数相同，用于 GBDT 模型；
- h) `LinearSVR_RFE_train_predict_rolling(dff, roll_begin_index = 51, para_search_step = 20, cv_score = 'neg_mean_absolute_error', cv = 5, random_state=300)`函数，用于线性 SVR 滚动预测，其中采用了 RFE 特征选择方法，如果换成 RFE 非线性核，暂时不支持特征选择；
- i) `RBF_SVR_train_predict_rolling(dff, roll_begin_index = 51, para_search_step = 20, cv_score = 'neg_mean_absolute_error', cv = 5, pcc = 0.1)`函数，SVR 中采用 RBF 核，特征选择采用皮尔逊相关系数，`pcc` 为 0.1，默认每次滑动选择 `pcc` 大于 0.1 的特征；

3. evaluation 模块：

- a) `evaluate_func(df, self = True, level = 0, threshold = 0)`函数：返回
MAPE,ACC,SEN 等，self 默认为 True，即预测值和实际值都是自己和自己比；
- b) `test_stationarity(timeseries, window=365)`函数：用于时间需的平稳性检测
- c) `trend_split(ts, freq = 52)`函数：用于时间序列的趋势分解
- d) `cal_wave_pcc(df, threshold_up = '75%',threshold_down = '25%')`函数：用于
计算波峰波谷相关性函数，threshold_up 为大于多少人为是波峰，
threshold_down 为小于多少为波谷，默认为上中和下中位数；

4. visualization 模块：

- a) `plot_feature_import(dfu)`函数：特征相关性排序图，dataframe 的第一列为 y
- b) `plot_diff_shift(df,diff_lag = 10,shift_lag = 2,title = u'发病人数在滞后与差分后相关性')`函数：时间序列滞后、差分后相关性图
- c) `plot_diff_zhihou_corr(dfx, dfy,diff = 2,lagmax = 10)`：时间序列在与特征滞后、差分后的相关性；
- d) `plot_correlation_map(df)`：相关性矩阵热图
- e) `plot_zhihou_corr(dfx, dfy, title,lag_max = 30,kind = 'line')`：时间序列在与特征滞后的相关性；
- f) `get_p_value(df1,df2,name)`：获取相关性 p 值
- g) `swx_scatter_matrix(frame, alpha=0.5, beta = 0.8)`函数：相关性及特征分布矩阵图，alpha 为透明度，beta 为控制 text 的背景颜色