基于深度学习模型的空气质量预测研究报告

中国人民大学 信息学院 赵义金 2018104106

1、项目目标

使用深度学习融合模型，根据时序空气污染物数据和气象数据进行 PM2.5 数值预测

2、数据集

2.1、创建过程

从国际空气质量监测站官网（http://aqicn.org/city/all）获取沈阳市 11 个站点的从 2016 年 10 月 30 日 0：00 点至 2017 年 10 月 30 日 24：00 的每小时的空气污染物数据和气象数据，数据量共计 10 万条。

2.2、数据格式

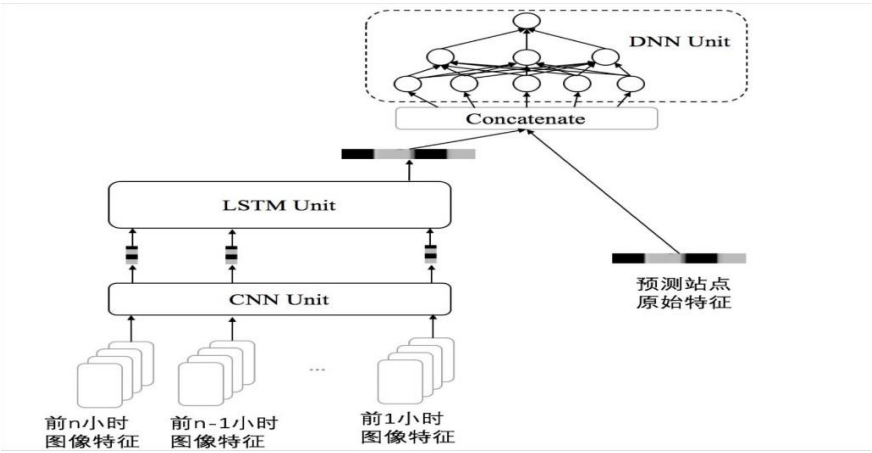
	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	mname	mtime	SO2	N0	NO2	NOx	O3	CO	PM10	PM2.5	pressure	temperatu	humidity	wind_E	wind_W	wind_N	wind_S	season	level
2	东陵路	2017/1/1 1:00	45	3	39	44	34	1.5	138	105	1015.1	-0.6	78	1.233856	0	0.40939	0	4	3
3	东陵路	2017/1/1 2:00	46	3	39	43	34	1.4	148	106	1015.1	-0.5	78	0	0.957143	0	1.154936	4	3
4	东陵路	2017/1/1 3:00	50	3	41	46	32	1.6	147	114	1014.9	-0.4	78	0	0.870414	0	1.342528	4	3
5	东陵路	2017/1/1 4:00	45	3	36	40	38	1.4	158	118	1014	-0.3	78	1.258622	0	1.142747	0	4	3
6	东陵路	2017/1/1 5:00	48	3	37	41	37	1.5	149	122	1013.9	-0.3	77	0.851758	0	0	1.471227	4	3
7	东陵路	2017/1/1 6:00	64	4	54	59	16	1.9	160	123	1013.5	-0.7	78	0.011967	0	0	1.499952	4	3
8	东陵路	2017/1/1 7:00	67	3	49	54	22	1.8	160	130	1013.6	-0.7	77	0.199826	0	1.48663	0	4	3
9	东陵路	2017/1/1 8:00	69	4	50	57	20	1.7	162	129	1014	-0.9	77	0	1.2557	0	0.820498	4	3
10	东陵路	2017/1/1 9:00	75	10	55	70	16	2	161	138	1014.7	-1.2	76	0	1.476093	0	0.266736	4	3
11	东陵路	2017/1/1 10:00	78	12	53	71	22	1.9	176	148	1015	-0.8	75	1.393201	0	0	0.555869	4	3

2.3、数据特性

数据为时间序列并且连续的；将真实的 PM2.5 值作为监督学习的标签。

3、模型介绍

3.1、体系结构



3.2、复杂度

3 层模型，每层对应的节点个数分别为 128，64，32 个。

3.3、模型特点

利用 CNN 卷积神经网络和 LSTM 对时序数据进行处理，将得到的向量通过全连接神经网络进行数值预测，使用真实 PM2.5 数值作为标签进行监督学习。

4、实验

4.1、输入输出

输入：

输出：

评价指标：$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{m}}$

4.2、训练开销

每个站点训练 1000 轮左右达到拟合。

4.3、性能分析

	MODEL	CNN	LSTM	ARIMA	SVM	GBDT
新秀街	9.563	12.371	11.448	12.784	10.872	10.977
文化路	10.781	13.773	11.489	12.005	12.973	11.734
小河沿	4.028	7.812	9.189	5.513	7.004	10.763
太原街	7.763	8.932	10.124	10.937	8.103	9.071
东陵路	6.218	10.277	9.173	7.249	6.992	10.372

时期	时间范围	污染源数量
Ⅰ-采暖期	2016 年 1-2 月	42
Ⅱ-非采暖期	2016 年 6-7 月	9
Ⅲ-采暖期	2017 年 1-2 月	67
Ⅳ-非采暖期	2017 年 6-7 月	17

预测时期	算法	PM2.5 浓度预测结果
Ⅰ-采暖期	CNN	10.871
	LSTM	9.672
	Arima	9.747
	SVM	10.287
	CNN-LSTM-全连接网络	9.019
Ⅱ-非采暖期	CNN	10.848
	LSTM	9.199
	Arima	10.443
	SVM	9.475
	CNN-LSTM-全连接网络	8.619
Ⅲ-采暖期	CNN	11.184
	LSTM	10.093

	Arima	10.884
	SVM	10.668
	CNN-LSTM-全连接网络	9.532
Ⅳ-非采暖期	CNN	10.754
	LSTM	10.002
	Arima	9.413
	SVM	9.874
	CNN-LSTM-全连接网络	8.467

五、总结

CNN-LSTM-全连接神经网络模型在 PM2.5 数值预测上比单独的 CNN、LSTM 以及统计学习模型例如 SVM 等有更好的效果；同时，将站点的地理位置信息信息考虑进去可以提升模型预测的准确度。