Getulio Vargas Foundation
Applied Mathematics School

Machine learning aproach for Dengue forecasting - Comparing LSTM, Random Forest and Lasso

Elisa Mussumeci

Rio de Janeiro
2018

Elisa Mussumeci

Using hierarchical clustering of timeseries for variable selection in Dengue forecasting

Dissertação submetida à Escola de Matemática Aplicada como requisito parcial para a obtenção do grau de Mestre em Modelagem Matemática da Informação.

Área de Concentração:

Orientador: Flávio Codeço Coelho

Rio de Janeiro
2018

Acknowledgements

Gostaria de agradecer....

Resumo

resumo...

Abstract

We use the Infodengue database of incidence and climate time-series, to train predictive models for the weekly number of cases of dengue in 733 cities of Brazil. To overcome limitation in the length of timeseries available to train the model, we included the time series of similar cities as predictors in the model of each city. The LSTM recurrent neural network model attained the highest performance in predicting future incidence on dengue in cities of different sizes.

Contents

1 Introduction **8**

2 Literature Review **10**

2.1 Epidemic forecasting 10

2.1.1 Multi time series forecast 10

2.1.2 Data modeling and cluster analysis 11

2.2 Neural networks 11

2.2.1 Long-short term memory (LSTM) 11

2.2.2 Neural networks and forecasting 11

3 Article **12**

3.1 Methodology 12

3.1.1 Data sources 12

3.1.2 Data modeling 12

3.1.3 Forecasting 13

3.2 Results 13

3.2.1 Cluster analysis 14

3.2.2 Forecasting 14
3.3 Discussion 14

4 Conclusions and Final Considerations **15**

References **16**

Appendices **16**

List of Figures

Chapter 1

Introduction

Understanding and therefore being able to predict the incidence of seasonal diseases is a big challenge due in part to the complex cycles these diseases display but also to incomplete records of historical disease incidence and other cofactors affecting risk. Besides the cycles are strongly influenced by local climate and other contextual variables making it hard to extrapolate findings from one geographical area to another.

For vector-borne diseases, the complexity is compounded by the coupling of the transmission of the dynamics in humans with the population dynamics of the vector species.

Having complete datasets for large geographical areas can help this effort as one can study the effects of spatial and climatic gradients on the intrinsic dynamics of disease transmission.

In this paper, we use the Infodengue[ref] database of more than 700 mu-

nicipalities of Brasil to develop predictive models capable of predicting the weekly incidence of Dengue in various regions of Brazil across a wide range of latitudes and climate characteristics.

Chapter 2

Literature Review

2.1 Epidemic forecasting

falar sobre o que é uma predição epidemica e quais metodos tem sido utilizados até então.

Explicar um pouco de cada método.

2.1.1 Multi time series forecast

Apresentar a diferença entre predição de series temporas simples e multiplas. Explicar cada modelo que vem sendo utilizado em prediçao multipla

2.1.2 Data modeling and cluster analysis

2.2 Neural networks

2.2.1 Long-short term memory (LSTM)

2.2.2 Neural networks and forecasting

Chapter 3

Article

3.1 Methodology

3.1.1 Data sources

For the forecasting model, we used data from the Infodengue project. Weekly incidence, minimum and maximum temperature, minimum and maximum humidity and atmospheric pressure series wher obtained for every city in the dataset.

3.1.2 Data modeling

Cities were clustered based on the correlation distance (eq xx) between incidence time series within each state.

For each city, a feature matrix was assembled from the set of time series

of each other time series from its cluster.

3.1.3 Forecasting

A LSTM model was defined with topology given in table xx. the model was trained for 300 epoch using a custom loss function defined in equation (xx) A look back of 4 weeks a forecasting window of 10 weeks were chosen.

A single city model was trained for a few selected cities to serve as a baseline against which to compare the effectiveness of the using sister cities (within the same cluster) cluster as predictors.

Random Forest

Long short term memory (LSTM)

Tpot - Lasso

3.2 Results

The cluster found within each state are shown in figures ... The clusters can also be seen in the map in figures xxx. Figures xx and yy show the performance of the prediction both *in-sample* and *out-of-sample*.

3.2.1 Cluster analysis

3.2.2 Forecasting

3.3 Discussion

The model has show good performance for both large and small cities from various parts of Brasil. This shows that the set of predictor series selected is capable to characterize the epidemic dynamics.

The extra information provided by the sister cities' series alowed the model to substantially outperform the base model. The LSTM model was capable of consistently predict the incidence pattern of non-epidemic years.

Chapter 4

Conclusions and Final Considerations

llalalala