Clustering of Covid-19 Time Series

Lisa Pucknat, Alexander Zorn

Lab Development and Application of Data Mining and Learning Systems:
Data Science and Big Data

9st March of 2021

UNIVERSITÄT BONN

Problem Definition

Task: Analysis of Covid-19 pandemic

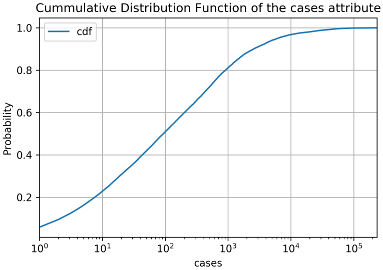
- Usage of dataset with daily Covid-19 cases
- Clustering algorithms for time-series to find clusters by countries and timespans
- Prediction of future cases using cluster analysis of the results

Framework

- We developed a flexible and modular Python framework consisting of 4 main parts:
  - Data generation and representation
  - (cluster) Model training and evaluation
    - using sklearn, sklearn-extra and tslearn
  - Prediction
  - Visualization
- Each step is fully managed using 2 configuration files
- multiple different configuration testing possible in one run

Data Analysis

- Dataset: multiple country daily Covid-19 case (and death) report set from the European Centre for Disease Prevention and Control
- Containing about 60.000 entries from 212 different countries from 12/31/2019 to 12/12/2020 and covering around 70 Mill. cases
- High share of zero $(31,3\%)$ and many low non zero case reports
- In total a very clean dataset
  - No twofold case reports.
  - Only about 470 missing intermediate reports
  - Only 17 negative reports



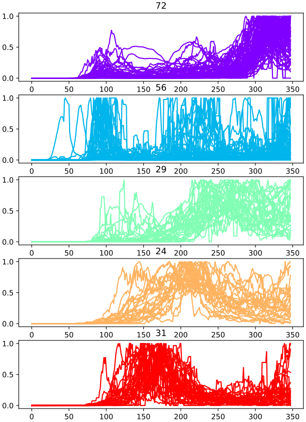
Clustering

Allocate countries with similar case developments into the same groups.

Challenges:

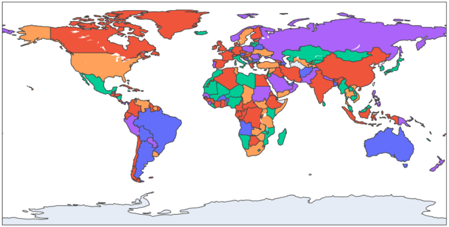
1. Shifts in time-series result in poor comparison → sim. measurement DTW
2. Unbalanced cluster distribution → Standardize values
3. Number of clusters → DBSCAN as a first assessment

Cluster Results



*KMeans, **ED**, K=5,*
avg. over 7 days

We used KMeans and KMedoids in combination with different similarity measurements, number of clusters, smoothing.



*KMeans, **DTW**, K=5,*
avg. over 7 days

Forecast

Predict next day with 30 preceding days available.

Two main approaches:

- Or train models on complete train data and use test data to forecast
- Use preceding snippet-clustering.
  Methods use the avg. cluster values or models trained only on data from the cluster

Experimented with two shallow methods:

- Naive- and Seasonal-Naive forecast

And two more complex methods:

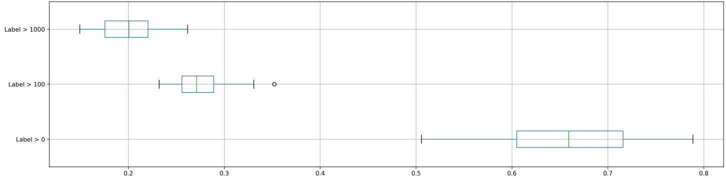
- fully connected linear neural network, LSTM network

Forecast Results

Experiments were conducted on 4 different 80/20 splits

- removing good or bad results by chance

The true forecast value is decisive for success of forecast



Naive forecast w/o cluster achieves prediction precision of 25% on avg.
Linear n.n. of 20% with best configuration[1]

[1] When filtering for labels above 1000 cases

Outlook

Developed modular and flexible framework, clustered Covid-19 cases by country and time-span and predicted the next day.

- Use ECDCs death values
  - more challenging, because of worse data situation
- Add more additional information such as temperature, age and health infrastructure
- Prediction of more than one value

We are happy to answer your questions!