Adaptive Drift Detection Adaptation for COVID-19 Forecasting

Universiteit Leiden
Nimish Pandey & Mehul Upase

3rd February 2025

1 Introduction and summary of the selected paper

The COVID-19 pandemic has created a great demand for public health forecasting in ways previously unimaginable and has put a premium on accurate predictions as guides to policy decisions. However, such forecasts are burdened by significant challenges due to the dynamic nature of the pandemic setting usually characterized by a concept drift, which is a phenomenon whereby changes occur in the distribution of data with time. The new variants, vaccination campaigns, and the change in public behavior are all factors that continuously shift the underlying patterns in COVID-19-related data, hence reducing the accuracy of any model, if drifts are not well addressed.

The original research paper, Automated Machine Learning for COVID-19 Forecasting [1], demonstrated the impact of AutoML on resource-efficient and accurate model forecasting. The authors applied multiple concept drift handling experiments: retraining, refitting, and partial refitting. Although these methods gave insights into managing drift, they relied on static strategies without significant real-time adaptability; hence, they may be quite inefficient and computationally expensive.

Our project therefore introduces, within the AutoML framework, an adaptive mechanism for the detection of drift. The mechanism, opposed to static strategies, will dynamically detect and react to drift in real-time. Anchoring on the statistical and machine learning-based methods for detecting drift, it selects the appropriate adaptation strategy based on the severity of the detected drift, from no action on negligible drift down to partial updates or full retraining.

These are two contributions are:

- An adaptive framework that includes the integration of drift detection along with dynamic response mechanisms to make the COVID-19 forecasting models more responsive and efficient.
- Perform a thorough evaluation of the proposed approach on various real-world COVID-19-related datasets regarding the examination of forecasting accuracy, computational efficiency, and drift detection performance.

The contribution of the current study lies beyond the static drift strategies, and the ambition is to establish a more robust and computationally efficient solution for handling concept drift in pandemic forecasting, thus contributing to a wider area of time-series forecasting under dynamic conditions.

2 Problem statement

COVID-19 forecasting faces significant challenges due to concept drift, when changes in the distribution of this data across time occur because of such factors as new variant emergences, vaccination rates, and public behavior. The current approaches are either computationally expensive, such as full model retrainings, or cannot answer well to the time-varying drift severity, such as partial refit. This project proposes an adaptive drift detection mechanism that will dynamically detect drift in real-time and take appropriate responses, from no action for negligible drift to partial updates or full retraining for acute drift. This adaptive approach will be aimed at improving the accuracy of the forecasts, reducing computational overhead, and enhancing the robustness of models to support shortcomings of fixed drift management strategies in evolving real-world scenarios.

So, we can describe the problem of COVID-19 forecasting under concept drift as a time-series forecasting problem that has dynamic data distribution changes, thus:

- $X = \{x_1, x_2, \ldots, x_t\}$ denote the historical time series data (i.e., mortality or mobility data) up to time t.
- $Y = \{y_1, y_2, \ldots, y_t\}$ denotes the target variable (i.e., mortality rates).
- w represent the input window size (no. of historical observations used for forecasting).
- h denote the forecasting horizon (no. of future steps to predict).

The goal is to forecast $\hat{Y} = \{y_{t+1}, y_{t+2}, \ldots, y_{t+h}\}$ using an input window $X_w = \{x_{t-w+1}, \ldots, x_t\}$. However, due to concept drift, the joint distribution $P(X, Y)$ changes a lot over time, which literally causes the models trained on historical data to degrade a lot in performance.

Concept drift occurs when:

$$P_{t_1}(X, Y) \neq P_{t_2}(X, Y), \quad \text{for } t_1, t_2 \in \mathbb{T} \text{ and } t_1 \neq t_2.$$

This drift can manifest as:

1. **Feature drift:** $P_{t_1}(X) \neq P_{t_2}(X)$
2. **Label drift:** $P_{t_1}(Y) \neq P_{t_2}(Y)$
3. **Conditional drift:** $P_{t_1}(Y|X) \neq P_{t_2}(Y|X)$

Let δ be a drift detection score that we will compute from the incoming data stream. So the adaptive mechanism would apply the following decision rule based on the severity of δ:

$$\text{Action} = \begin{cases} \text{No Update}, & \text{if } \delta < \epsilon_1 \\ \text{Partial Refit}, & \text{if } \epsilon_1 \leq \delta < \epsilon_2 \\ \text{Full Retrain}, & \text{if } \delta \geq \epsilon_2 \end{cases}$$

where ϵ_1 and ϵ_2 are drift severity thresholds.

Now, clearly, our objective is to basically minimize the forecasting error while efficiently handling drift:

$$\min_{\theta} \mathbb{E}\left[L(Y, \hat{Y})\right], \quad \text{subject to computational constraints,}$$

where: - $L(Y, \hat{Y})$ is the loss function (e.g., RMSE, MAE). - θ represents the model parameters optimized dynamically based on drift detection.

As, per the above formulations it can capture the problems of forecasting under dynamic data distributions, and also the need for adaptive tweaking to maintain model performance

3 Research Questions

1. How effectively can an adaptive drift detection mechanism identify and respond to concept drift in COVID-19 data?
2. How does tailoring drift responses improve forecasting accuracy and reduce computational overhead?
3. How robust is the adaptive mechanism across different drift scenarios (gradual, sudden, recurring) using real-world datasets?

4 Methodology

In this project to address the challenges of concept drift in COVID-19 forecasting through the development of an adaptive mechanism for drift detection along with AutoML pipelines. The journey starts with the task of data preparation, where publicly available COVID-19 mortality and mobility data are extracted from sources like WHO and Google Mobility Reports. Normalization of data is carried out on these metrics to ensure consistency among the different regions, whereas missing values are imputed using techniques like moving averages. These datasets range from training, validation, and test sets that can be used to simulate some real-world scenarios of concept drift, including gradual, sudden, and recurring changes.

It is a methodology with a focus on the mechanism of drift detection so that changes in data distribution can be observed. To this end, statistical tests such as the Kolmogorov-Smirnov test and Jensen-Shannon divergence will be applied, as well as online detection methods like ADWIN and DDM (Yet to be finalized). It constantly contrasts recent data with historical distributions using a sliding window approach. Thus, it could find out the exact instance of drift and quantify its magnitude in real-time.

The mechanism of adaptive response would handle concept drift dynamically, depending on the detected severity: negligible drift-continue with the current model, no updates; mild drift-partial refit on recent data; and severe drift-full retraining or replacement of the model. The thresholds for the severity of the drift are set using metrics computed from the outputs of detections so that the responses can be tailored to the nature of the drift.

This is done by incorporating the adaptive mechanism in the already existing AutoML frameworks like auto-learn. The extension of AutoML pipelines by adding the preprocessing step of drift detection within every pipeline, and also the incorporation of adaptive response strategies regarding model updates. Lightweight configurations are used so that it can be computationally efficient, which allows it to scale and be responsive to real-time changes in data.

This methodology is based on the use of dynamic detection and response mechanisms in order to enhance the robustness and efficiency of the forecasting models and provide a better fit for the evolution characteristics of COVID-19 data. The method proposed here differs from traditional strategies since it is capable of adapting dynamically to drift.

5 Evaluation approach

The proposed adaptive drift detection will be evaluated from a number of dimensions. Forecasting accuracy: this will be captured through metrics such as RMSE and MAE, benchmarking the performance of the adaptive mechanism relative to traditional drift management strategies (e.g., full retrain and partial refit) and baseline models (e.g., ARIMA, deep learning).

To assess the performance of the algorithms on drift detection, we will use the metrics: precision, recall, and F1-score. These metrics will tell us how effective the algorithms under consideration are at finding the expected concept drift and issuing necessary responses. We will also be doing an analysis of computational efficiency: the run time and resource utilization including memory and CPU usage of each setup will be measured. This is done by studying different concept drift scenarios- sudden, gradual, and recurring drifts-on real-world COVID-19 datasets. This approach surely will ensure that the most comprehensive assessment of the framework's accuracy, efficiency, and robustness is covered.

Ethical statement

- Privacy Concerns: Data for COVID-19 forecasting is aggregate and anonymized data on mortality, mobility, and vaccination rates coming from reputable organizations such as WHO and Google Cloud. No personal identifiers or sensitive individual-level information are included; hence, privacy risks are minimal.
- Model Bias: The model may adopt some of the biases from the data feeding into the model, such as disparities in COVID-19 testing, reporting accuracy, and health access at varying levels across regions and demographics. For instance, regions with limited testing report fewer cases, hence skewing predictions.

It will be trained from data across many countries and regions; much care will be taken in its validation process so that large discrepancies in performance do not occur as a result of this diversity.

- Forecasting Data Misuse Risk: The model-generated forecasts could either be misinterpreted or used out of context. Improper application of predictive data in decision-making relating to public health, for instance, leads to panic, complacency, or poorly conceived policy measures. The results should be clearly and transparently presented by describing the limitations of forecasts and quantify uncertainties associated with them; this would be a key component of ensuring that policymakers and the public understand the model is generating probability forecasts, not deterministic predictions.
- Community Impact: Accurate forecasting aims to help communities by informing public health responses, yet incorrect predictions could potentially lead to negative outcomes, such as unnecessary restrictions or insufficient precautions. Thorough validation and conservative prediction ranges will be employed to minimize such risks.

Overall, the project emphasizes transparency, rigorous evaluation, and responsible communication of results to ensure ethical and equitable use of the forecasting model.

Division of workload

- Nimish: Implement statistical and online drift detection techniques; set severity thresholds, Design dynamic responses (no update, partial refit, full retrain) for drift severity,
- Mehul: Collect, preprocess, and normalize datasets; simulate drift scenarios, Extend AutoML pipelines to include drift detection and adaptive updates, Visualize results and compare with baselines.
- Both will Test adaptive mechanisms on diverse drift scenarios and collect results, and work equally on presentation and report.

Code:

- **Link to the Repository where our code will be uploaded:**
  https://github.com/noshamedevil/UC-Final-Project
- **Link to the Code (Which will be used as a base of work for this project):**
  https://github.com/AutoML4covid19/Forecasting/tree/main

Datasets to be used:

Data Sources: Mortality and mobility datasets from WHO and Google Cloud's COVID-19 open data. All datasets are publicly available.

Ethical Considerations: Data is anonymous, with no personal identifiers; ethical risks are minimal.

- https://data.who.int/dashboards/covid19/data
- https://www.ecdc.europa.eu/en/publications-data/sources-worldwide-data-covid-19
- https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country
- https://health.google.com/covid-19/open-data/raw-data

References

[1] Jaco Tetteroo, Mitra Baratchi and Holger H. Hoos. "Automated Machine Learning for COVID-19 Forecasting". In: *IEEE Access* 10 (2022), pp. 94718–94737. DOI: 10.1109/ACCESS.2022.3202220.