THE INTERNATIONAL JOURNAL OF BUSINESS & MANAGEMENT

Comparative Analysis and Application of Deep Neural Networks in Covid-19 Prediction

Dome Rogers
Ph.D. Student, Masinde Muliro University of Science and Technology, Kakamega, Kenya

Abstract:
Machine learning specifically deep or reinforcement learning is a growing field. With the advent of the SARS COV 2 pandemic, its application for Covid 19 prediction has continued to evolve as traditional prediction approaches are rendered inaccurate or outdated with the challenges paused as we globally try to understand the trend of this global pandemic. Deep learning coupled with bio-statistical theory and approaches provide a new way of tackling the prediction challenge in Covid 19.This new approach factors in complex variation in variables that is non-linear, multivariate and with multiple independent variables. We assess the promise entailed in automated machine learning, SIR & Hybrid SIR Models referred to as SEIR (Susceptible-Exposed-Infected-Recovered) Models and LSTM RNN (Long-Short Term Recurrent Neural Networks. These three approaches directly inform CDC, WHO, major entities sharing prediction results on the pandemic and individual government health organs globally. We explore why they exhibit efficiency in arriving at predictions as the variables, geography and demographics fed into each keep varying. The criticality of the assessment we arrive at is rigorously tested and validated using K-Fold validation, Mean Accuracy Prediction Error (MAPE) and we also plot receiver operating characteristic (ROC) curves the results are later on exhibited showing Auto-ML.

Keywords: *AutoML, Hybrid SIR, SIR Model, SEIR Model, LSTM RNN, Covid 19 prediction*

1. Introduction

Data as a currency in the current century continues to gain importance, and it's focal in the Covid 19 response and in ultimately arriving at a durable solution the world over. There are many public and private data aggregation and generation entities serving as both traditional and non-traditional data actors and stakeholders. These bodies unilaterally agree that one of the main challenges is gaps in the data to aid in response and understanding the Covid 19 virus both biological and socio-economic effects.

Currently these gaps can be categorized as follows:

- Covid 19 gender data disaggregation gaps. Thought the main Covid 19 data is disaggregated by gender on deaths, infected and recovered, we have few cases exposing the pandemics deep rooted impact in granular socio-economic data and effect on gender beyond the aforementioned categories of aggregation.[1]This limit predictive modelling on binary data male-female supervised learning on trend predictions like regression analysis are strongly inhibited.
- Comorbidities data gaps. Covid 19 being a relatively new and understudied disease, the data available is limited. However, from the cases that emerged, it was observed that comorbidities increase the chances of infection. This shows the vital role data on Covid 19 comorbidities will play and shines a light on this data gap.[2]This limit supervised learning like time to event analysis with independent variables as comorbidities are strongly inhibited.
- With Covid 19 impacting the globe exponentially, the race factor is becoming key in accurate death rate modelling[3] and how we respond to the pandemic. There are race and ethnicity data gaps that continue to limit this[4].

There are multiple machine learning algorithms implemented differently by the various tools and applications in predicting trends traditionally used in supervised learning. The most reliable non-traditional techniques owing to the fact that we have multiple algorithms, may lead to the application of the automated machine learning approaches. Either of these techniques are crucial because of the volume of data and the need to identify algorithms with the best accuracy. This is a classic prediction machine learning problem. The use of automated machine learning[5] to automate model identification

[1]'COVID-19: Gendered Impacts of the Pandemic in Palestine ….' 6 May. 2020, https://www.un.org/unispal/document/covid-19-gendered-impacts-of-the-pandemic-in-palestine-and-implications-for-policy-and-programming-un-women-analysis/.
[2]'Comorbidity and its Impact on Patients with COVID-19.' 25 Jun. 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7314621/.
[3]'Racial Disparities in COVID-19: Key Findings from Available ….' 17 Aug. 2020, https://www.kff.org/report-section/racial-disparities-in-covid-19-key-findings-from-available-data-and-analysis-issue-brief/.
[4]'COVID-19 race and ethnicity data reporting still a challenge in ….'https://www.inquirer.com/health/coronavirus/coronavirus-covid-19-race-data-reporting-20200809.html.
[5]'(PDF) Automated Machine Learning: State-of-The-Art and ….' 5 Jun. 2019, https://www.researchgate.net/publication/333650038_Automated_Machine_Learning_State-of-The-Art_and_Open_Challenges.

across 1000's of models, to test, validate and identify for production the most effective and accurate model to use, introduces the need for a rapid iteration in ML model building and production.

1.1. Model Selection and Application

The CDC,WHO,John's Hopkins, Kenya MOH and KNBS are a wealth of knowledge in availing resources, data and studies on supervised learning techniques that inform our approaches.

- SIR Compartmental Regression Model, expressed as SIR (Susceptible-Infected-Recovered). This Models is traditionally designed around the epidemiological approach informed by SIR (Susceptible-Infected-Recovered) modelling approach[6].We have traditional SIR Models (IHME is used to illustrate this) and the Hybrid SIR (SEIR) Model, The Google and Harvard Model is used to explain this.
- LSTM RNN. Ensemble models as opposed to singular models are superior[7] and more accurate by reviewing daily prediction data over the last 3 months and comparing death rate prediction models from more than 30 modelling groups/think tanks[8].We showcase IHME's LSTM ensemble model. This is more advanced than SIR.
- AutoML (Automated Machine Learning)[9].Since social distancing levels vary and are constantly adjusted, factoring these into the models affects accuracy, and increases it extremely. We showcase AutoML to build Covid 19[10]Model. Various algorithms are auto-selected, auto-tuned and accuracy presented for the number of pipelines (recursive model selection and tuning)

1.2. SIR and Hybrid SIR

For SIR and Hybrid SIR which involve the modelling of the susceptible, infected and recovered, while hybrid SIR involved modelling of the three inclusive of exposure rate and tightening and loosening of lockdowns and curfews, are modeled on Covid19 secondary statistics as follows:
Infection rate, Positivity rate, Tested population. R0 number of people a single infected person can infect (2 to 3 people), Recovery rate (10 to 14 days).

1.3. LSTM RNN

Normalization technique was applied using keras as indicated in the code snippet below, whereby we made the data row wise to add up to 1. Inverse scaling was also used.

1.4. Auto ML

- Various owing to the approach, which ensures a random set of ML techniques is selected and pipelines containing each is run several times each run called a generation for the length of time allocated on google TPU (tensorflow processing unit using google colab), I ran the tensors for 240 Minutes iterating across three different models.
- All the selected models are inherently supervised regression model variants with the ability to handle time series data.

All the pipelines utilized in the AutoML techniques above had very low accuracy scores, after having validated using K fold validation technique whereby we had, test, training and validation data broken into 5 validation data sets, to prevent overfitting.

- The SIR and Hybrid SIR is a compartmental model - we model and predict the susceptible population, the infected and recovered. For the hybrid SIR we include modelling and prediction of the exposed. We take into recognition that people progress through these stages. We use multivariate time-series regression, which falls under supervised machine learning.
- LSTM RNN - This is a deep learning, neural network that is recurrent, meaning it has both forward and backward feed, commonly known as long-short term memory recurrent neural network it's a combination of an LSTM and RNN.We ran 10 epochs.
- AutoML - automating the process of applying machine learning. We use a set of approaches whose performance we then evaluate (feature importance, model performance and others). Since we have clearly defined metrics on the individual steps, we can automate the process. Here we get to the AutoML search for optimal solutions across methods, features, transformations and parameter values.

Model performance:

- SIR and Hybrid SIR we compared predicted to actuals using test and training sets of data, with time series data changing daily a plot of these two curves illustrated the difference clearer.
- LSTM RNN - We split the data into training and testing, we had a validation set that we broke into multiple (K fold validation fold) and we performed kfold validation. In this case our K was 5.

[6]'COVID-19 dynamics with SIR model · The First Cry of Atom.' 11 Mar. 2020, https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html. Accessed 3 Sep. 2020.
[7]'ensemble - CDC.' 13 Jul. 2020, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/pdf/Consolidated-Forecasts-2020-07-13.pdf. Accessed 3 Sep. 2020.
[8]'COVID-19 Forecasts: Cases | CDC.' 29 Aug. 2020, https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html. Accessed 3 Sep. 2020.
[9]'A fully automatic deep learning system for COVID-19 ....'https://erj.ersjournals.com/content/56/2/2000775. Accessed 7 Sep. 2020.
[10]'A Fully Automatic Deep Learning System for COVID-19 ....' 16 May. 2020, https://erj.ersjournals.com/content/early/2020/05/19/13993003.00775-2020. Accessed 7 Sep. 2020.

- AutoML - We split the data into training and testing, we had a validation set that we broke into multiple (K fold validation fold) and we performed kfold validation. In this case our K was also 5 as above.

1.5. SIR and Hybrid SIR

The accuracy of our model is by comparing the MAPE (mean absolute percentage errors - this is a percentage of absolute errors) results, the lowest is the most accurate from a select SIR models, our model is second from right, blanks are from dates I couldn't find reliable data online for.

1.6. LSTM RNN

The accuracy for this was the highest and this was the most reliable model, though scores as high as 90% could be an overfitting of the model for future predictions we tested future predictive power and validated that it works for 7 days into the future see below:

1.7. Auto ML

This had the poorest accuracy though it's supposed to be the most superior.

1.8. Receiver Operating Characteristic (ROC) area

ROC For LSTM RNN indicates high sensitivity, this indicates strong predictive power of the model.
The ROC for the AutoML model is least accurate and close to zero showing none sensitivity.
SIR Model ROC Curve, this had the second-best sensitivity further from zero.

2. Conclusion and Recommendations

LSTM RNN has the best performance and we can also confirm predictive accuracy would last for how many days into the future. I highly recommend this even though AutoML allows automation of the process across multiple models, the best model is the LSTM RNN. Though a lot of work has been done on SIR and Hybrid SIR, and compartmental models are key in epidemiological studies of this nature, with regard to the data we chose, Johns Hopkins, the LSTM RNN is most accurate, with lowest MAPE of 5% and accuracy of 94% consistently.

- Hybrid SIR and LSTM RNN can complement each other, by introducing compartments in the recurrent neural nets.
- AutoML brings the power of testing and creating pipelines across multiple pipelines, this can be explored with optimization from algorithm categories that are accurate on predictive studies for diseases.
- Introduction of modelling that assess changes in lockdown intensity increases accuracy since the time variant is lagged.

3. List of Acronyms

AutoML	- Automated Machine Learning
BI-LSTM	-Bi-Directional Long-Short Term Memory
CDC	-Centers for Disease Control and Prevention
CNN	- Convolutional neural network
ICU	-Intensive Care Unit
IHME	-Institute for Health Metrics and Evaluation
K Fold	-5-Fold
Kenya MOH	-Kenya Ministry of Health
KNBS	-Kenya National Bureau of Statistics
LSTM RNN	-Long-Short Term Memory Recurrent Neural Network
MAE	-Mean Absolute Errors
MIT	-Massachusetts Institute of Technology
ML	-Machine Learning
R0	-Number of people an individual can infect
RMSE	-Root Mean Square Error
SEIR	-Susceptible-Exposed-Infected-Recovered
SIR	-Susceptible-Infected-Recovered
TPOT	-Python Automate Machine Learning Tool
UCLA	-University of California, Los Angeles
UK	-United Kingdom
US	-United States
WHO	-World Health Organization
YYG	-Youyang Gu

4. References

i. COVID-19: Gendered Impacts of the Pandemic in Palestine ….' 6 May. 2020, https://www.un.org/unispal/document/covid-19-gendered-impacts-of-the-pandemic-in-palestine-and-implications-for-policy-and-programming-un-women-analysis/.
ii. Comorbidity and its Impact on Patients with COVID-19.' 25 Jun. 2020,

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7314621/.
iii. Racial Disparities in COVID-19: Key Findings from Available ….' 17 Aug. 2020, https://www.kff.org/report-section/racial-disparities-in-covid-19-key-findings-from-available-data-and-analysis-issue-brief/.
iv. COVID-19 race and ethnicity data reporting still a challenge in ….' https://www.inquirer.com/health/coronavirus/coronavirus-covid-19-race-data-reporting-20200809.html.
v. Automated Machine Learning: State-of-The-Art and ….' 5 Jun. 2019, https://www.researchgate.net/publication/333650038_Automated_Machine_Learning_State-of-The-Art_and_Open_Challenges.
vi. COVID-19 dynamics with SIR model · The First Cry of Atom.' 11 Mar. 2020, https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html.
vii. Ensemble - CDC. 13 Jul. 2020, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/pdf/Consolidated-Forecasts-2020-07-13.pdf.
viii. COVID-19 Forecasts: Cases | CDC.' 29 Aug. 2020, https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html.
ix. A fully automatic deep learning system for COVID-19 ….' https://erj.ersjournals.com/content/56/2/2000775.
x. A Fully Automatic Deep Learning System for COVID-19 ….' 16 May. 2020, https://erj.ersjournals.com/content/early/2020/05/19/13993003.00775-2020.
xi. Comparative Study of Best Time-Series Models for Urgent ….' 26 Aug. 2020, https://www.datasciencecentral.com/profiles/blogs/comparative-study-of-best-time-series-models-for-urgent-pandemic.
xii. COVID-19 Human Data Exchange (HdX) data source https://covid19.healthdata.org/.
xiii. COVID-19 Forecasts: Deaths | CDC. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html.