Disease Outbreaks: Tuning Predictive Machine Learning

Tassallah Abdullahi, Geoff Nitschke
abdtas008@myuct.ac.za,gnitschke@cs.uct.ac.za
Department of Computer Science
University of Cape Town, South Africa

ACM Reference Format:
Tassallah Abdullahi, Geoff Nitschke. 2021. Disease Outbreaks: Tuning Predictive Machine Learning. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459414

EXTENDED ABSTRACT

Climate change is expected to exacerbate diarrhoea outbreaks in developing nations, a leading cause of morbidity and mortality in such regions [3]. The development of predictive models with the ability to capture complex relationships between climate factors and diarrhoea may be effective for diarrhoea outbreak control. Various supervised *Machine Learning* (ML) algorithms and *Deep Learning* (DL) methods have been used in developing predictive models for various diseases [7]. Despite their advances in a range of healthcare applications, overall method task performance still largely depends on available training data and parameter settings which is a significant challenge for most predictive machine learning methods [9]. This study investigates the impact of *Relevance Estimation and Value Calibration* (REVAC) [4], an evolutionary parameter optimization method applied to predictive task performance of various ML and DL methods applied to ranges of real-world and synthetic data-sets (diarrhoea and climate based) for daily diarrhoea outbreak prediction in a regional case-study (South African provinces). Preliminary results indicate that REVAC is better suited for the DL models regardless of the data-set used for making predictions.

Data-sets

This study's data-sets focused on nine South African Provinces: *Western Cape*, *Eastern Cape*, *Northern Cape*, *North West*, *Free State*, *Limpopo*, *KwaZulu Natal*, *Gauteng*, and *Mpumalanga*. For each province, a ten year period (2008−2018) of daily sales records of *Loperamide*, an anti-diarrhoea compound evaluated in the treatment of patients with chronic non-specific diarrhoea in South Africa was obtained from *Clicks* pharmaceuticals. These data were used as proxy for diarrhoea cases in the region. The number of diarrhoea cases per day for a specific province was computed as the number of Loperamide sales per day in the given province. Data-sets also contained an 11 year period (2008−2019) of six-hourly data on eight climate factors[1] for each province attained from the centres for *Atmospheric Research* and *Atmospheric Prediction*[2].

Generative Adversarial Networks (GANs) [1] were used to generate 20,000 synthetic time-series samples for the diarrhoea data and eight climate data in each country province. This synthesis was performed to have sufficient data for making predictions. Augmentation with considerable amounts of synthetic data improves performance while augmentation with too many synthetic data inhibits task performance [1]. Results indicated augmentation with 20, 000 synthetic data-objects to complement the real data worked best in our experiments. Synthetic data was augmented with real-world data using *upward* and *downward* augmentation. For data-sets augmented *upwards*, training set included combinations of real-world and synthetic data, but test sets included only synthetic data. For *downwards* augmented data-sets, training sets included mainly synthetic data and test sets included real-world data.

Machine Learning (ML) Algorithms

Three ML algorithms were evaluated for predicting daily diarrhoea cases: *Support Vector Machines* (SVMs) [6], and two DL algorithms: *Long Short-Term Memory* (LSTM) and *Convolutional Neural Networks* (CNNs) [2]. We used *Python Scikit-Learn*[3] for support vector regression to develop all our SVMs with a *Radial Basis Function* (RBF) Kernel [6]. DL models were designed with the *Keras DL Library and TensorFlow*[2]. For all ML methods, we used REVAC to automate parameter tuning [4]. SVM method parameters tuned were γ and C, and *dropout rate*, *pool size*, *neurons*, *batch size*, *learning rate*, *epochs*, *filter size* and *layer size* were tuned for DL methods (otherwise we used default *Keras* DL package values). Before method training and testings, all data-sets were normalized and divided into a ratio of 70:30 for training and testing. To evaluate predictive task-performance of each method, we used the *Root Mean Square error* (RMSE, figure 1), based on previous research recommendations [5].

Relevance Estimation and Value Calibration

REVAC automates meta-heuristics (originally evolutionary algorithm) parameter tuning [4], where given an objective, task environment, parameter vector population and N iterations, REVAC uses evolutionary optimization to explore, select, and evaluate sets of beneficial parameter values. Each iteration, REVAC improves and updates the distribution of the parameter vectors until an optimal or near-optimal meta-heuristic task performance is attained. Here, REVAC was implemented as a meta-layer that aided in searching for optimal parameter values for each of our ML methods (SVM, LSTM, CNN) when applied to the task of predicting diarrhoea cases.

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459414

[1] Climate variables included: *Maximum and Minimum temperature, Air temperature, Humidity, Evaporation and Precipitation rate, Surface pressure, Wind velocity.*
[2] https://psl.noaa.gov/ | https://keras.io/
[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

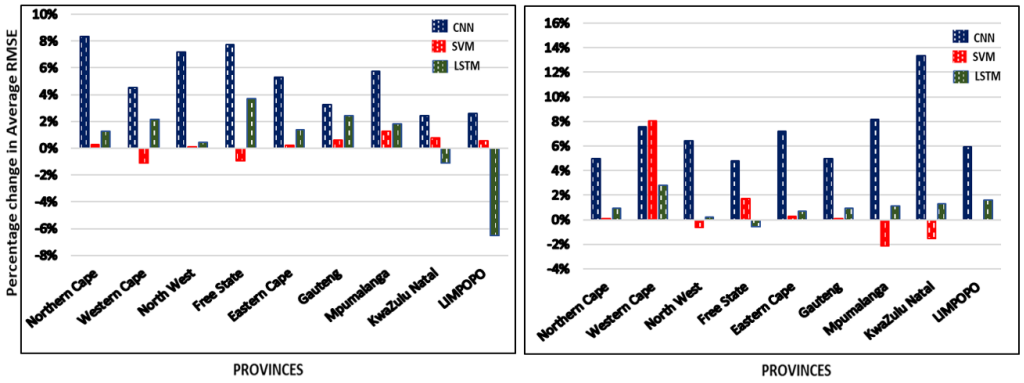

Figure 1: Percentage change in performance of each ML model when predictions were made with parameters from REVAC tuning instead of the *grid-search* parameters for (left) *upward* augmented and (right) *downward* augmented data-set. High percentages indicate an improve in task performance.

Experiments

Experiments[4] investigated the impact of REVAC parameter tuning on ML method task performance for predicting daily number of diarrhoea cases. We compared average RMSE of the three methods when *grid-search* [9] was used to tune parameter values versus REVAC parameter tuning given varying portions of augmented data used as input. Our experiments used parameter tuning methods to select combinations of possible parameter values from user-specified value ranges. The parameters tuned with both REVAC and Grid-search are listed in ML Algorithms section.

In the first experiment, grid-search was used to find suitable parameter values for each ML method across each country province, using varying portions of both *upward* and *downward* augmented data as input. For each province and each data-set the optimal parameter values of each ML method selected by grid-search was used to make final method predictions. In the second experiment, REVAC was used to find optimal parameter values for each ML method across each province. The same portions of *upward* and *downward* augmented data for each province was also used as input for tuning each ML method. The objective of the REVAC was to minimize a fitness function which was RMSE predictions made by each ML method for specific input data. All REVAC parameters are from related work [4]. The best REVAC parameters for each method and province were used for final method predictions.

Results and Discussion

Results indicate that when real-world data is augmented with more than 50% of synthetic data, task performance of all ML methods starts to decline regardless of parameter tuning, where final results were computed by averaging RMSE of all augmented data-set portions. Figure 1 presents results indicating the task performance of the CNN model improved significantly over each province (Wilcoxon test, $p<0.05$) for all data-sets. Comparatively, the prediction accuracy for the LSTM also increased for some data-sets (provinces), but there was a performance drop in predictive accuracy for other provinces (data-sets) including: *Limpopo*, *KwaZulu Natal* and *Free State*. Wilcoxon tests ($p<0.05$) indicated that these declines were not significant, indicating that REVAC parameter tuning may still be appropriate for LSTM methods. However, SVM task performance declined for Western Cape, KwaZulu Natal, *Free State*, North West and Mpumalanga provinces (data-sets). Though the decline in performance for these provinces were not significant, the provinces where SVM task performance improved were not statistically significant either. We surmise that REVAC parameter tuning is not ideal for SVM methods but rather better suited to the tuned DL method parameters given such noisy, incomplete and augmented data-sets and predictive tasks, as presented in this study.

This is hypothesized to be a result of REVAC's improved efficacy on high-dimensional search spaces [4], as well as demonstrated effectiveness of CNNs with augmented data-sets [2, 8]. Also, related work [9] indicated that grid-search is better suited to low dimensional search spaces as constituted by SVM method parameters (SVM methods using RBF kernels need only *gamma* and *C* parameters [6]). Current research is investigating evolutionary parameter tuning methods that potentially automate parameter tuning of any given predictive ML method applied to sparse, noisy and data-augmented data-sets, to thus design auto-tuning methods applicable to predicting a vast range of disease outbreak tasks.

[4]https://github.com/ProjectRepo2021/REVAC-tuning-outbreak-prediction

REFERENCES

[1] Z. Che and et al. Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. In *Proceedings of International Conference on Data Mining*, pages 787–792. IEEE, 2017.
[2] Y. Le Cun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(1):436–444, 2015.
[3] G. Musengimana and et al. Temperature Variability and Occurrence of Diarrhoea in Children under Five-years-old in Cape Town Metropolitan Sub-districts. *International Journal of Environmental Research and Public Health*, 13(9):859, 2016.
[4] V. Nannen and A. Eiben. Relevance estimation and value calibration of evolutionary algorithm parameters. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 975–980. Morgan Kaufmann, 2007.
[5] R. Pelanek. Metrics for Evaluation of Student Models. *Journal of Educational Data Mining*, 7(2):1–19, 2015.
[6] N. Sapankevych and R. Sankar. Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38, 2009.
[7] D Sathya, V Sudha, and D Jagadeesan. Application of Machine Learning Techniques in Healthcare. In *Handbook of Research on Applications and Implementations of Machine Learning Techniques*, pages 289–306. IGI, 2020.
[8] L. Taylor and G. Nitschke. Improving Deep Learning with Generic Data Augmentation. In *Proceedings of the IEEE Symposium Series on Computational Intelligence*, pages 1542–1547. IEEE, 2018.
[9] J. Wu and et al. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, 17(1):26–40, 2019.