Representation Learning to Integrate Dynamic and Static Data for Improved Prediction on COVID-19 Patients Mortality

Anonymous ICCV submission

Paper ID ****

Abstract

Seasonal influxes of COVID-19 infections put pressure on healthcare systems and make the allocation of limited resources and treatments extremely difficult. Because clinical measurements of COVID-19 patients can be collected over time, several longitudinal deep learning models have been proposed to predict clinical outcomes and aid in logistical decision making. These models often fail, however, to effectively handle missing data or integrate static data with the multivariate time series (MTS).We propose a novel semi-supervised learning framework that leverages Long Short-Term Memory (LSTM) to learn the vectorial representation of MTS which can be easily integrated with the static data. Using this representation, learned from both labeled and unlabeled samples, one can fully utilize the information in the MTS dataset. Experimental results on chest X-ray images dataset show that the proposed model significantly improves the mortality classification compared to the supervised classifiers.

1. Introduction

1. Which problem we are going to solve? 2. Why solving this problem is important? 3. Which studies have beed tried to solve this problem and what are the disadvantages/difficulties of them? 3-1. Which properties of longitudinal image datasets should be considered to solve the problem? (e.g. time series, inconsistent number of scans for different participants, some are labeled others are not labeled, ..) 4. So which model we are going to propose to solve the listed problems above? 4-1. Learn enriched representation of longitudinal images data in a fixed-length vector format which can be readily integrated with the static data. 4-2. Convolutional layer + LSTM to enrich the images. 4-3. The proposed model is Semi-supervised learning model. [4]

Sudden increases in COVID-19 cases are incredibly stressful on healthcare systems and quickly deplete the all ready limited resources of hospitals [2]. Clinicians are forced to make difficult descisons about the distribution of scarce treatments without reliable predictions about patients' conditions. To aid in this problem, machine learning models have been proposed to inform logistical planning such as in the study by [8]. This model[8] identifies the most predictive biomarkers for a patient's mortality using a XGBoost classifier [1] trained on a publicly available MTS dataset collected from COVID-19-positive patients admitted to Tongji Hospital in China. Due to the limitations of random XGBoost classifier in synthesizing longitudinal data, only the final record was used to train and test the model. While the model effectively determines the most predictive biomarkers, it hardly captures the temporal trends, which are critical in determining the progression of the disease. It also fails to predict the mortality when those principle biomarkers are not measured. An intrinsic problem of the COVID-19 pandemic is that samples from patients are rarely collected in a uniform way. X-rays of patients are not preformed for a controlled study, resulting in a different number for every patient as well as irregular time intervals between samples. MTS analysis models have been proposed to learn relationships across variables and time stamps, however, conventional MTS analysis often requires sequences to be complete and of consistent length [3, 7, 6, 5] which is problematic for the reasons stated above. Furthermore, datasets collected out of carefully controlled environments tend to have missing data which is particularly problematic when that includes patient outcome.

2. Methods

Methods here.

3. Experiment

Experiment here.

4. Conclusion

Conclusion here.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 1

[2] Centers for Disease Control, Prevention (CDC), et al. Strategies for optimizing the supply of n95 respirators. 2020 apri 3 https://www. cdc. gov/coronavirus/2019-ncov/hcp/respiratorsstrategy/index. html. *CDC_AA_refVal= https% 3A% 2F% 2Fwww. cdc. gov% 2Fcoro navirus% 2F2019-ncov% 2Fhcp% 2Frespiratorsstrategy% 2Fcrisis-alternate-strategies*. 1

[3] Lyujian Lu, Hua Wang, Xiaohui Yao, Shannon Risacher, Andrew Saykin, and Li Shen. Predicting progressions of cognitive outcomes via high-order multi-modal multi-task feature learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 545–548. IEEE, 2018. 1

[4] Kerem Sinan Tuncel and Mustafa Gokce Baydogan. Autoregressive forests for multivariate time series modeling. *Pattern recognition*, 73:202–215, 2018. 1

[5] Hua Wang, Feiping Nie, Heng Huang, Jingwen Yan, Sungeun Kim, Shannon Risacher, Andrew Saykin, and Li Shen. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *Advances in neural information processing systems*, pages 1277–1285, 2012. 1

[6] Xiaoqian Wang, Dinggang Shen, and Heng Huang. Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–281. Springer, 2016. 1

[7] Xiaoqian Wang, Jingwen Yan, Xiaohui Yao, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, Heng Huang, et al. Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model. In *International Conference on Research in Computational Molecular Biology*, pages 287–302. Springer, 2017. 1

[8] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for covid-19 patients. *Nature Machine Intelligence*, pages 1–6, 2020. 1