

DOCTORAL THESIS

Deep Learning Models for Irregular Electronic Health Record Data Analysis

TAN, Qingxiong

Date of Award:
2021

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

The wide adoption of electronic health records (EHR) produces a large quantity and variety types of health data, such as laboratory parameters, disease diagnoses, treatment reports, medication orders, as well as some basic health data (e.g., birth date, gender), which provides unprecedented opportunities for developing advanced machine learning models to detect high-risk patients, evaluate early treatment effects, and assist doctors to design more effective treatment plans, effectively improving healthcare quality and clinical outcomes.

However, the analysis of EHR data is challenging because real-world medical data is highly irregular (significantly varying time intervals between successive visits and severe missing data problem in the examinations of physiological variables) and heterogeneous (from stable basic health data to significantly changing laboratory parameters). First, influenced by the dynamic changes in the severity of illness, patients usually go to hospitals and take examinations irregularly, producing a large volume of irregular medical time-series data with varying intervals. Meanwhile, certain physiological variables are not examined during some visits to the hospitals, causing the missing data problem. Furthermore, there could be some relationships between different types of medical data, e.g., basic health data may influence future trajectories of dynamic variables. To handle these issues, in this thesis, a series of novel deep learning-based approaches are proposed for EHR data analysis.

In this thesis, we develop an uncertainty-aware neural network to deal with the irregular visit intervals. Existing methods often handle this problem by generating regular time series from the irregular medical records without considering the

uncertainty in the generated data, induced by the varying intervals. Thus, a novel Uncertainty-Aware Convolutional Recurrent Neural Network (UA-CRNN) is proposed in this thesis, which introduces the uncertainty information in the generated data to boost the risk prediction. To tackle the complex medical time series with sub-series of different frequencies, the uncertainty information is further incorporated into the sub-series level rather than the whole sequence to seamlessly adjust different time intervals. Specifically, a hierarchical uncertainty-aware decomposition layer (UADL) is designed to adaptively decompose time series into different sub-series and assign them proper weights in accordance with their reliabilities. Meanwhile, an Explainable Uncertainty-Aware Convolutional Recurrent Neural Network (eUA-CRNN) is proposed to exploit filters with different passbands to ensure the unity of components in each sub-series and the diversity of components in different sub-series. Furthermore, eUA-CRNN incorporates with an uncertainty-aware attention module to learn attention weights from the uncertainty information, providing explainable prediction results. Extensive experimental results on three real-world medical datasets illustrate the superiority of the proposed method compared with the state-of-the-art methods.

On the other hand, varying time intervals and missing values may provide valuable information in improving clinical prediction performance because visit intervals are usually determined by the health status of patients while missing values are caused by changes in symptoms of patients. As a result, the traditional strategy of processing irregular EHR data into regular data to fit the properties of standard machine learning models may damage the meaningful medical considerations in the irregular examination arrangements. Therefore, we develop a novel end-to-end Dual-Attention Time-Aware Gated Recurrent Unit (DATA-GRU) for irregular EHR data analysis to avoid damaging the informative varying intervals and missing data phenomenon. In particular, DATA-GRU is able to: 1) preserve the informative varying intervals by introducing a time-aware structure to directly adjust the influence of the previous status in coordination with the elapsed time, and 2) tackle missing values

by proposing a novel dual-attention structure to jointly consider data-quality and medical-knowledge. A novel unreliability-aware attention mechanism is designed to handle the diversity in the reliability of different data, while a new symptom-aware attention mechanism is proposed to extract medical reasons from original clinical records. Extensive experimental results on two real-world datasets demonstrate that DATA-GRU can significantly outperform the state-of-the-art methods and provide meaningful clinical interpretations.

Furthermore, since multiple data sources (e.g. dynamic laboratory parameters, static basic health data) can reflect different aspects of the health conditions of patients, designing a framework to effectively combine them for clinical predictions will potentially improve the performances. To this end, in this thesis, we develop a novel End-to-end Importance-Aware Personalized Deep Learning Approach (eiPDLA) to achieve accurate early clinical risk prediction. Specifically, eiPDLA introduces a Long Short-Term Memory with Temporal Attention to learn sequential dependencies from time-stamped records and simultaneously incorporating a Residual Network with Correlation Attention to capture their influencing relationship with static medical data. Furthermore, a new Multi-Residual Multi-Scale Network with the importance-aware mechanism is designed to adaptively fuse the learned multi-source features, automatically assigning larger weights to important features while weakening the influence of less important features. Extensive experimental results under various settings on real-world medical data illustrate that the proposed method significantly outperforms the state-of-the-art methods. Case studies indicate that the achieved prediction results are highly interpretable.

The main contributions of this thesis are summarized as follows.

- We propose a novel explainable uncertainty-aware neural network to incorporate the uncertainty information in the generated data to dynamically learn contribution weights of different data and introduce filters to adaptively adjust the unity of components in each sub-series and the diversity of components between different sub-series. Meanwhile, an attention module is incorporated

to identify key features and provide explainable prediction results.

- We propose a novel dual-attention time-aware neural network to jointly deal with the two key challenges (i.e., varying time intervals and missing values) in irregular EHR data by preserving the contained useful information on dynamic changes in the health status of patients.
- We propose a novel end-to-end importance-aware approach to simultaneously model multiple data sources, including both static and dynamic health data, while adaptively learning fusion weights for different deep features and assign larger weights to important features.

Keywords: irregular EHR data analysis, uncertainty-aware prediction, dual-attention mechanism, explainable prediction results, multiple data sources

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	vi
Table of Contents	viii
List of Tables	xii
List of Figures	xiv
Chapter 1 Introduction	1
1.1 Background	1
1.1.1 Electronic Health Record Data	1
1.1.2 Challenges in EHR Data Analysis	3
1.2 Motivation of This Project	6
1.3 Dissertation Focus and Contributions	9
1.4 Thesis Overview	12
Chapter 2 Related Work	14
2.1 EHR Data Mining Methods	14
2.1.1 Modeling of Single Type of EHR Data	14
2.1.2 Joint Modeling of Multiple Types of EHR Data	16
2.2 Uncertainty-Aware Mechanism	16

2.3	Attention Mechanism	17
-----	-------------------------------	----

Chapter 3	Explainable Uncertainty-Aware Convolutional Recurrent Neural Network for Irregular Medical Time Series	19
3.1	Introduction	19
3.2	Proposed Method	23
3.2.1	Gaussian Processes for Data Processing	25
3.2.2	Uncertainty-Aware Decomposition Layer	26
3.2.3	Explainable Uncertainty-Aware Decomposition Layer	29
3.2.4	Residual Network Convolutional Layers	36
3.2.5	Gated Recurrent Unit Recurrent Layers	36
3.2.6	Objective Function	37
3.3	Experiments	38
3.3.1	Dataset Description	38
3.3.2	Experimental Settings	39
3.3.3	Comparing Methods	40
3.3.4	Implementation Details	42
3.3.5	Experimental Results Analysis	43
3.3.6	Explainability and Sub-series Analysis	50
3.4	Conclusion	54
3.5	Experimental Results on More Types of Input Features	54
3.6	Experimental Results on Different Confidence Intervals	55

Chapter 4	DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series	57
4.1	Introduction	57
4.2	Proposed Method	61
4.2.1	Notations	62
4.2.2	Time-Aware Mechanism	63
4.2.3	Dual-Attention Mechanism	65

4.2.4	Interpretable Embedding	67
4.2.5	Objective Function	67
4.3	Experiments	68
4.3.1	Data Description and Experimental Settings	68
4.3.2	Comparing Methods	68
4.3.3	Results and Discussion	69
4.3.4	Case Study	73
4.4	Conclusion	75

Chapter 5 Importance-Aware Personalized Learning for Early Risk Prediction Using Static and Dynamic Health Data 76

5.1	Background and Significance	76
5.2	Proposed Method	81
5.2.1	Correlation Analysis	82
5.2.2	Temporal Analysis	85
5.2.3	Deep Feature Fusion	88
5.2.4	End-to-End Importance-Aware Personalized Deep Learning	90
5.3	Experiments	93
5.3.1	Dataset Description	93
5.3.2	Experimental Settings and Evaluation Metrics	93
5.3.3	Comparing Methods	94
5.3.4	Results for Different Observation Windows	95
5.3.5	Results for Different Prediction Windows	98
5.3.6	Ablation Study	99
5.3.7	Discussion	101
5.4	Conclusion	105

Chapter 6 Conclusions 106

6.1	Summary	106
6.2	Future Work	108

6.2.1	Special Shapes Detections	108
6.2.2	Time-Frequency Analysis	109
	Bibliography	110
	List of Publications	137
	Curriculum Vitae	139

List of Tables

3.1	Comparison to the state-of-the-art methods on long-term mortality risk predictions for the PUB dataset, evaluated via AUC scores (mean \pm std).	43
3.2	Comparison to the state-of-the-art methods on short-term mortality risk predictions for the MIMIC-III dataset, evaluated via AUC scores (mean \pm std).	46
3.3	Comparison to the state-of-the-art methods on in-hospital mortality risk prediction with different observation windows for the eICU dataset, evaluated via AUC scores (mean \pm std).	48
3.4	The AUC scores of our proposed methods for different confidence intervals (CIs) on the MIMIC-III dataset (mean \pm std).	56
4.1	The AUC scores (mean \pm std) of different levels of mortality risk predictions for the MIMIC-III dataset.	70
4.2	The AUC scores (mean \pm std) of different levels of mortality risk predictions for the eICU dataset.	71
4.3	Patient information in the case study. All the timestamps use admission time of the patient as the benchmark.	73
5.1	The AUC scores of different observation windows for 6-month ahead mortality risk predictions (mean and 95% confidence interval).	96
5.2	The AUC scores of different observation windows for 12-month ahead mortality risk predictions (mean and 95% confidence interval).	97

5.3	The AUC scores of different levels of mortality risk predictions with 9-year observation window (mean and 95% confidence interval). . . .	99
-----	--	----

List of Figures

1.1	Illustration of irregular EHR time-series data.	4
1.2	Illustration of why the problems of missing data and the issues of irregular visit intervals are common in real-world EHR data.	5
1.3	Relationship between different proposed methods. Each proposed method will be described in detail in the corresponding chapters. . . .	10
3.1	An example of utilizing Gaussian Process (GP) to generate regular time series from irregular records. (a) The generated regular data fit well with observed records, but they may deviate from actual data in areas with sparse observation; (b) The estimated uncertainty evaluates unreliability of generated data by quantifying the degree in which they may deviate from actual data.	21
3.2	An overview of the proposed method for irregular medical time series. The architecture consists of four parts: 1) Gaussian Process (GP) for data imputation, which generates regular time series while estimating corresponding uncertainty information; 2) Uncertainty-aware decomposition layer (UADL), which takes the generated time series and the uncertainty information as inputs, decomposes time series, and assigns different weights (calculated from the uncertainty information by using decay function) to sub-series according to their stabilities; 3) ResNet convolutional layers, which learn sequence features from uncertainty-aware sub-series; 4) GRU recurrent layers, which generate risk scores from the sequence features.	24

3.3	Illustration of the proposed explainable uncertainty-aware decomposition layer (eUADL). Compared with UADL, eUADL has two significant improvements: 1) Uncertainty-aware attention module, which can identify key variables from different inputs and provide explainable risk prediction results; 2) Residual learning based filters, which ensure the unity of the components in each sub-series and the diversity the components between different sub-series, achieving better prediction performance.	30
3.4	The setting of the observation window and the prediction window. . .	39
3.5	The network configuration of UA-CRNN.	42
3.6	The ROC curves of long-term mortality risk prediction results for the PUB dataset.	44
3.7	A case study for explainability and sub-series analyses in the PUB dataset. (a) Weights directly transformed from uncertainty scores via decay function; (b) Weights learned by the attention module; (c) Sub-series of different frequencies obtained by eUADL; (d) Scores contributed by different decomposition channels.	51
3.8	A case study for explainability and sub-series analyses in the MIMIC-III dataset. (a) Weights directly transformed from uncertainty scores via decay function; (b) Weights learned by the attention module; (c) Sub-series of different frequencies obtained by eUADL; (d) Scores contributed by different decomposition channels.	52
3.9	A case study for explainability and sub-series analyses in the eICU dataset. (a) Weights directly transformed from uncertainty scores via decay function; (b) Weights learned by the attention module; (c) Sub-series of different frequencies obtained by eUADL; (d) Scores contributed by different decomposition channels.	53

3.10	Influence of the number of input features on the degree that eUA-CRNN outperforms UA-CRNN for different risk prediction tasks in the MIMIC-III dataset.	55
4.1	Illustration of irregular multivariate time series (IMTS). The time intervals between successive visits vary significantly from each other. Meanwhile, certain physiological variables are not examined at some visits, causing the missing value problem.	58
4.2	The architecture of the proposed DATA-GRU method and the illustration of the dual-attention mechanism. (a) DATA-GRU takes input records \mathbf{x}_t , time intervals Δ_t and unreliability scores \mathbf{u}_t as inputs. Parts in purple, green, and cyan denote the time-aware mechanism, symptom-aware attention, and unreliability-aware attention, respectively. DATA-GRU deals with irregular visit intervals by converting them into weights via a decay function to adjust the influence of the previous status on the current status. (b) DATA-GRU tackles the missing value problem by designing a novel dual-attention structure to jointly consider data-quality and medical-knowledge.	61
4.3	Visualization of attention weights. (a1) the weight matrix learned by unreliability-aware attention of DATA-GRU, which shows score assigned to each variable at each visit; (a2) the average weight of different variables at each visit; (b1) and (b2) provide weight matrix and average weight learned by symptom-aware attention of DATA-GRU.	72
4.4	Attention weights assigned to different variables.	74
5.1	Illustration of dynamic changes in laboratory parameters of two patients with different immune abilities. (a) A patient with weak immune ability, whose risk level is rather high in the end; (b) A patient with strong immune ability, whose status finally returns healthy. . . .	79

5.2	An overview of the proposed Personalized Deep Learning Approach (PDLA) for early prediction of clinical risk.	82
5.3	The architecture of eiPDLA and the illustration of the importance-aware mechanism.	90
5.4	Changes of AUC scores along the increase of observation window length for two different prediction windows (6 months and 12 months).	98
5.5	Ablation study on 12-month ahead mortality risk predictions.	100
5.6	Visualization of attention weights. (a) the attention weights learned for static health data; (b) the attention weights learned for dynamic health data; (c) time-series data of CR and UREA; (d) Contribution scores of CR and UREA.	103