KDD2017

Patient Subtyping via Time-Aware LSTM Networks

Inci M. Baytas [1] Cao Xiao [2] Xi Zhang [3] Fei Wang [3] Anil K. Jain [1] Jiayu Zhou [1]

[1]Computer Science, Michigan State University

[2]Center for Computational Health, IBM T.J. Watson,

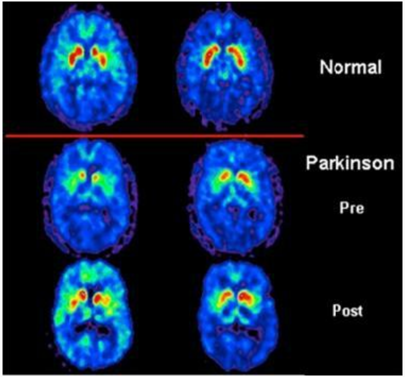
[3]Health Care Policy and Research, Weill Cornell Medical School

Overview

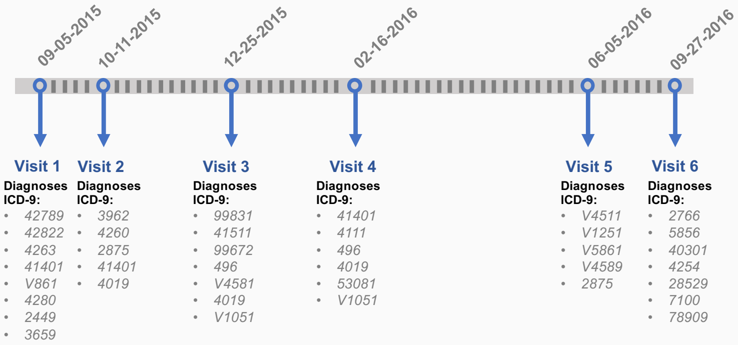
Introduction

Patient Subtyping

- Patient subtyping seeks patient groups with similar disease progression pathways based on longitudinal medical records.
- Subtyping is one of the common tasks to observe different types of conditions in Parkison's Disease (PD) studies.

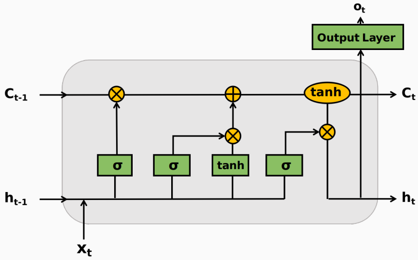


Electronic Health Records



- Clinical information of a patient is recorded at each visit or admission; Electronic Health Records (EHR).

Irregular Elapsed Time

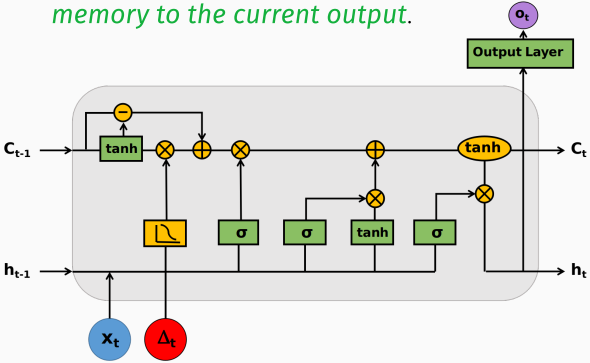


- One of the appealing approaches for sequence modeling is Long-Short Term Memory (LSTM).
- LSTM assumes the elapsed time is uniform throughout the sequence.
- It is very typical for health records to have time intervals varying days to years.
- Elapsed time has a significance in clinical decision making.
- It is crucial to capture the relationships and the dependencies between each record of a patient under time interval irregularities.

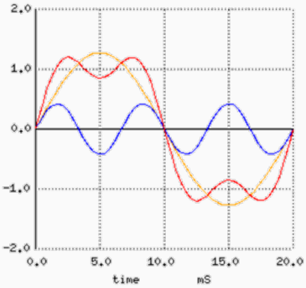
Time-Aware LSTM (T-LSTM)

T-LSTM Unit

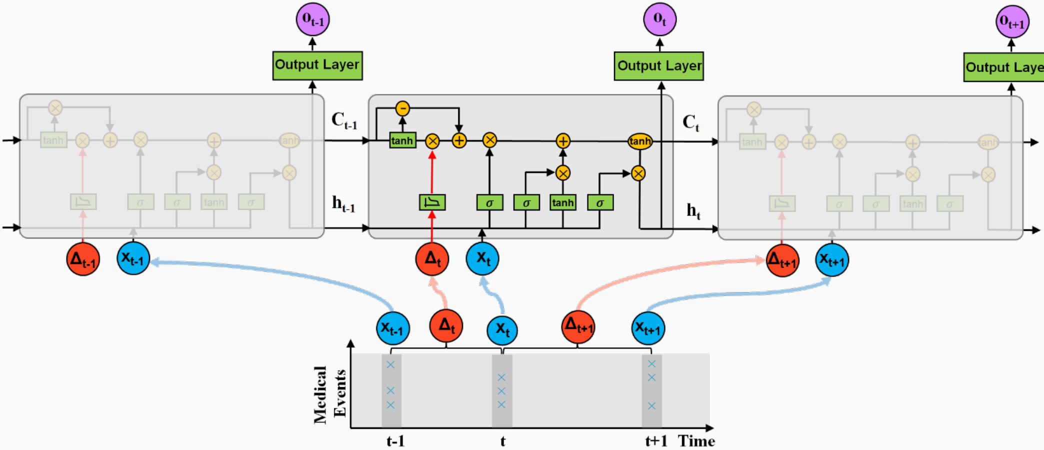
- T-LSTM incorporates the elapsed time such that *longer the elapsed time, smaller the effect of the previous memory to the current output.*



- We can think of subspace decomposition as decomposing mixture of sine waves.



T-LSTM in Healthcare

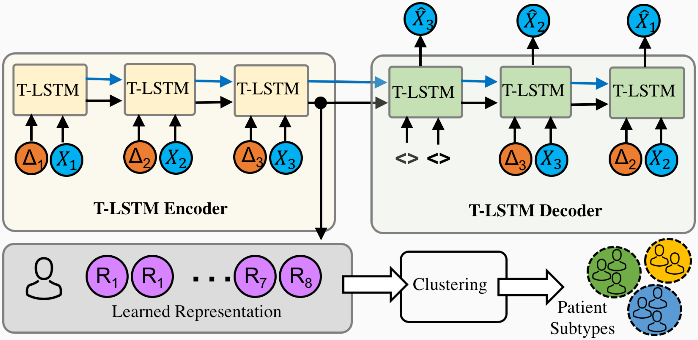


Patient Subtyping via T-LSTM

An Unsupervised Problem

- Patient subtyping is posed as a clustering problem since we do not have any prior information about the groups inside the patient cohort.
- An efficient representation summarizing the structure of the temporal records of a patient is required.
- Auto-encoders provide an unsupervised way to directly learn a mapping from the original data.
- T-LSTM auto-encoder was used to learn a mapping from the original patient sequence to a single representation.

T-LSTM Auto-encoder



After learning representations, patients can be grouped by using a clustering algorithm, such as k-means.

Experiments

Synthetic Data

- T-LSTM was tested for a supervised task on an artificially generated EHR data [1].
- Data has electronic records of up to 100,000 patients with lab results, diagnoses, and start and end dates of the admissions.
- Each patient has a unique patient ID similar to real world EHR data.
- Target diagnoses was Diabetes Mellitus;
  - **Task:** binary classification
  - **Input:** sequence of admissions
  - **Output:** predicted one-hot label
- 6,730 patients were sampled with an average of 4 admissions.
- Feature dimensionality was 529 (529 unique diagnoses present in the dataset).

[1] http://www.emrbots.org/

Results

Supervised synthetic EHR experimental results, average AUC of testing on 10 different splits. Training and testing ratio was chosen as 70% and 30%, respectively. MF1-LSTM and MF2-LSTM are two baselines where forget gate is modified by the elapsed time [2].

Methods	Avg. Test AUC	Stdev.
T-LSTM	**0.91**	0.01
MF1-LSTM	0.87	0.02
MF2-LSTM	0.82	0.09
LSTM	0.85	0.02
LR	0.56	0.01

[2] Pham et. al, DeepCare: A Deep Dynamic Memory Model for Predictive Medicine, arxiv:1602.00357v1 [stat.ML], 2016.

Parkinson's Progression Markers Initiative (PPMI) Data

- PPMI [3] is an observational clinical and longitudinal study comprising of evaluations of people with Parkinson's disease and people at high risk and healthy controls.
- Preprocessed PPMI [4] data was used in the experiments.
  - Number of patients: 654
  - Elapsed time interval: $[1, 26]$ months
  - Average sequence length: 25
  - Input feature dimensionality: 319
  - Target dimensionality: 82

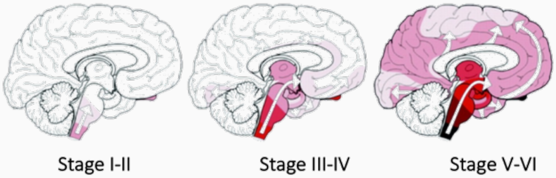
---

[3] www.ppmi-info.org

[4] Che et.al. " An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease.", SDM 2017.

PPMI Data Features

- Features include:
  - Demographics
  - Motor severity measures such as Unified Parkinson's Disease Rating Scale, Hoehn and Yahr staging,
  - Non-motor manifestations: depression, anxiety, cognitive status, sleep disorders,
  - Imaging assessment such as DaTScan.

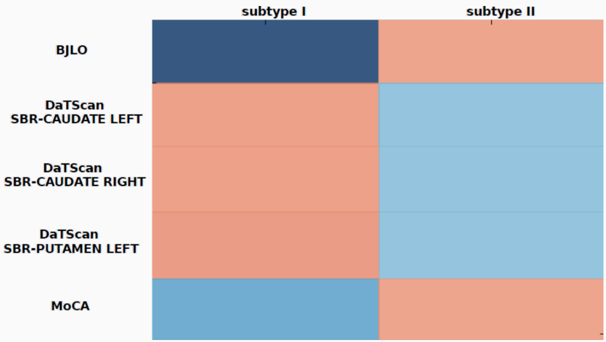


Statistical Analysis Results

Feature	P-Value	Cluster1 Mean	Cluster2 Mean
T-LSTM			
BJLO	9.51×10^{-8}	16.5	24.7
MoCA	0.001	40.0	41.2
DaTScan1	0.042	2.29	2.07
DaTScan2	0.027	2.31	2.08
DaTScan4	0.001	1.4	1.1
MF1-LSTM			
CSF-Total tau	0.007	87.9	46.72
MoCA	2.16×10^{-17}	47.5	41.05
SDM	0.005	58.5	41.5
MF2-LSTM			
HVLT-Retention	0.03	0.84	0.83
SDM	0.007	36.61	41.68

- Chi-square test for the categorical features, F-test for the normal continuous features, Kruskal-Wallis test for non-normal continuous features, and Fisher's exact test for sparse features.
- If the p-value is less than 0.05, a significant group effect is considered for the associated feature.
- Method producing more features with $p < 0.05$, is considered as providing a more sensible patient subtyping result.

Heatmap for T-LSTM



- Orange color represents the cluster mean which is higher than the total mean of the patients and the shades of blue show lower mean values for the corresponding feature with $p < 0.05$.
- PD patients are known to have lower DaTScan SBR values than healthy subjects.

Conclusion and Future Work

Summary

- A new LSTM unit which can deal with irregular elapsed times between the consecutive elements of sequential data is proposed.
- Time irregularity can be encountered in other domains, such as video sequence with missing frames.
- T-LSTM does not have any assumption about the elapsed time measure.
  - Different types of non-increasing function of elapsed time can be used for different time measures.
- In the future work, T-LSTM will be improved;
  - Different ways of incorporating the elapsed time
  - Interpretability

Thank You!