Adaptive Skip-Train Structured Regression for Temporal Networks

Martin Pavlovski[1,2], **Fang Zhou**[1], **Ivan Stojkovic**[1,3], **Ljupco Kocarev**[2], **Zoran Obradovic**[1]

[1] Computer & Information Sciences Department, Temple University
[2] Macedonian Academy of Sciences and Arts, Skopje, Macedonia
[3] School of Electrical Engineering, University of Belgrade, Belgrade, Serbia

Abstract

A broad range of high impact applications involve learning a predictive model in a temporal network environment. In weather forecasting, predicting effectiveness of treatments, outcomes in healthcare and in many other domains, networks are often large, while intervals between consecutive time moments are brief. Therefore, models are required to forecast in a more scalable and efficient way, without compromising accuracy. The Gaussian Conditional Random Field (GCRF) is a widely used graphical model for performing structured regression on networks. However, GCRF is not applicable to large networks and it cannot capture different network substructures since it considers the entire network while learning. In this study, we present the Adaptive Skip-Train Structured Ensemble (AST-SE), a sampling-based structured regression ensemble for prediction on top of temporal networks. Capable of automatically skipping the entire or some phases of the training process, AST-SE outperforms its competitors, while learning in a more efficient, scalable, and potentially more accurate manner.

Introduction

- **Problem statement:**
  - A network $G^{(t)} = (V^{(t)}, E^{(t)}, \mathbf{X}^{(t)}, \mathbf{y}^{(t)})$ is observed over time.
  - Objective: Given an unobserved network $G^{(t+1)} = (V^{(t+1)}, E^{(t+1)}, \mathbf{X}^{(t+1)})$, predict the response variable at each node $\mathbf{y}^{(t+1)}$
- **Graphical Models:**
  - Commonly used to predict the response at each node in one or multiple upcoming time steps.
  - Retrained at each step
- **Challenges:**
  - Time for prediction is limited
  - High Computational and space complexity
- **Goal:** Forecast in a more scalable and efficient way, without compromising accuracy.

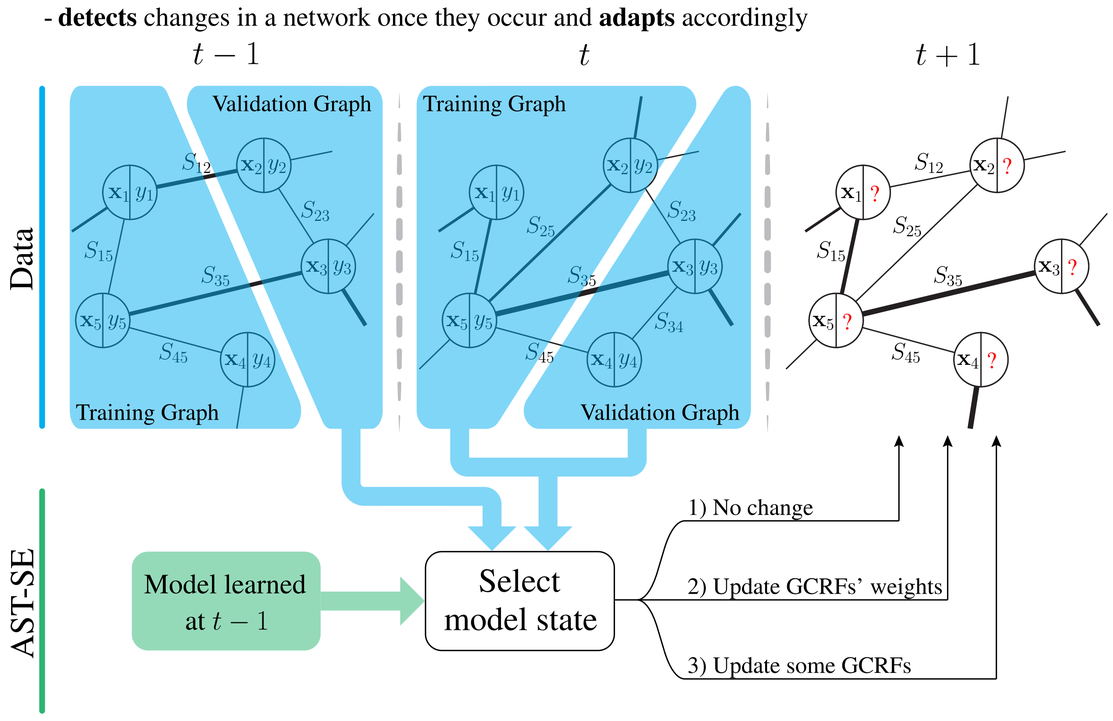
Gaussian Conditional Random Fields

A GCRF [1,2] models the conditional distribution:

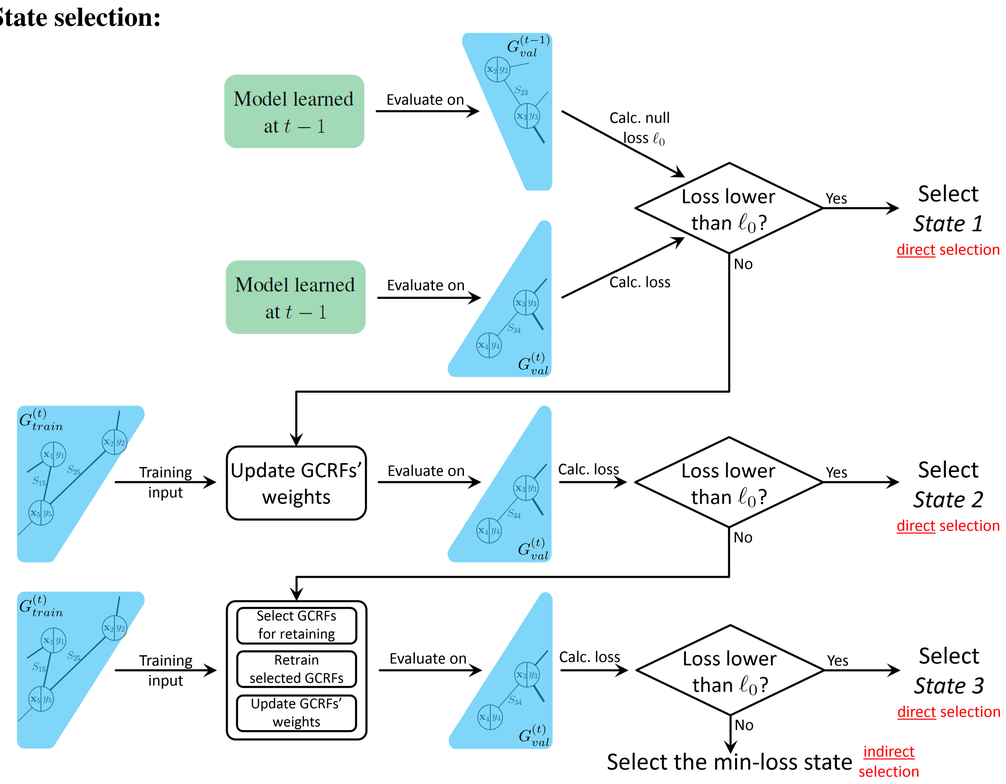
$$P(\mathbf{y}^{(t)}|\mathbf{X}^{(t)}) = \frac{1}{Z(\mathbf{X}^{(t)}, \alpha^{(t)}, \beta^{(t)})} \exp\left\{-\alpha^{(t)} \sum_{i=1}^{N} \left(y_i^{(t)} - R_i(\mathbf{X}^{(t)})\right)^2 - \beta^{(t)} \sum_{i \sim j} S_{ij}^{(t)} \left(y_i^{(t)} - y_j^{(t)}\right)^2\right\}.$$

Methodology

- To predict the outputs for all nodes at time step $t+1$, one can train a single GCRF or even a GCRF ensemble model at time step t.
- Repetitive retraining at each time step can be redundant as data distributions are often similar in consecutive time steps.
- To overcome this issue, AST-SE:
  - employs **multiple graphical models** in order to learn different relationships using network substructures
  - **detects** changes in a network once they occur and **adapts** accordingly



State selection:



AST-SE states:

- *State 1 (no change)*

$$\Phi_1^{(t)}(G^{(t+1)}) = \sum_{m=1}^{M} \omega_m^{(t-1)} \phi_m^{(t-1)}(G^{(t+1)}),$$

- *State 2 (reweight)*

$$\Phi_2^{(t)}(G^{(t+1)}) = \sum_{m=1}^{M} \omega_m^{(t)} \phi_m^{(t-1)}(G^{(t+1)}),$$

- *State 3 (retrain + reweight)*

$$\Phi_3^{(t)}(G^{(t+1)}) = \sum_{m=1}^{M^*-1} \omega_m^{(t)} \phi_m^{(t)}(G^{(t+1)}) + \sum_{m=M^*}^{M} \omega_m^{(t)} \phi_{s_m}^{(t-1)}(G^{(t+1)}).$$

Results

Experiments were performed on:

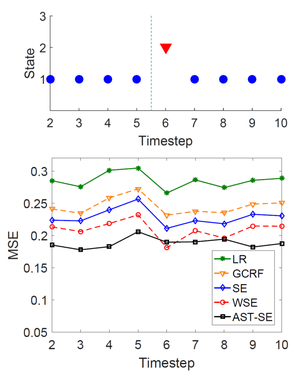
1) synthetically generated temporal networks
2) gene expression network [3] - a real-world temporal network

Experiments on synthetic temporal networks

- **Nodes:** 10,000 input-output pairs
- **Structure:** generated using an Erdős-Rényi random graph model
- **Task:** predict the output values at the next time step

Timestep	GCRF parameters	
	α	β
$t = 1, \ldots, 5$	1	[2,5]
$t = 6, \ldots, 10$	[2,5]	1

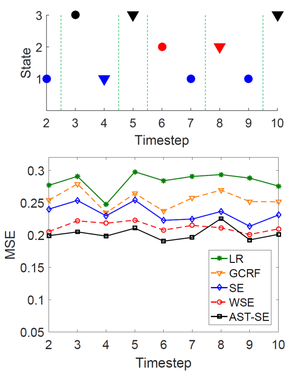
Model	Execution time
GCRF	5014.65 ± 360.43
SE	188.30 ± 9.03
WSE	207.52 ± 3.84
AST-SE	15.38 ± 28.06



#1: one data distribution change

Timestep	GCRF parameters	
	α	β
$t = 1, 2$	1	[2,5]
$t = 3, 4$	1	[6,9]
$t = 5$	1	1
$t = 6, 7$	[2,5]	1
$t = 8, 9$	[6,9]	1
$t = 10$	1	1

Model	Execution time
GCRF	5254.84 ± 259.37
SE	193.30 ± 11.14
WSE	217.14 ± 7.12
AST-SE	70.16 ± 34.49

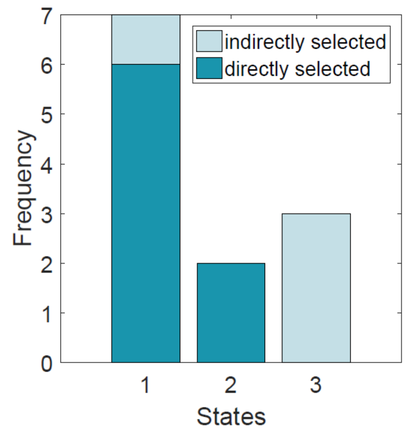


#2: data distribution changes more frequently

Real-World Application: Influenza Virus Network Prediction

- **Data:** Infuenza A virus subtype H3N2 network observed over time (16 hours/steps)
- **Nodes:** 12,032 genes
- **Features:** expression values from 3 previous time steps
- **Targets:** expression values at the current time step
- **Structure:** similarities between gene expressions
- **Task:** predict the expression values at the next time step

Model	MSE	Execution time
LR	0.38 ± 0.19	$\mathbf{0.10 \pm 0.03}$
GCRF	0.39 ± 0.21	9082.71 ± 1898.43
SE	0.39 ± 0.21	297.29 ± 19.42
WSE	0.35 ± 0.19	309.32 ± 19.44
AST-SE	$\mathbf{0.23 \pm 0.07}$	64.00 ± 45.73



Conclusions

- **Efficiency:** AST-SE is ∼**140** and ∼**4.5** times faster than GCRF and ensemble-based alternatives, respectively, when its components are run in parallel on the H3N2 Virus Influenza network.
- **Scalability:** AST-SE focuses only on partial views of a network, hence it is scalable as the network size expands.
- **Accuracy:** AST-SE obtains a ∼**34-41%** smaller average error (MSE) when compared against alternatives on the H3N2 network.

Acknowledgments

This research was supported in part by DARPA grant No. FA9550-12-1-0406 negotiated by AFOSR, the National Science Foundation grants NSF-SES-1447670, NSF-IIS-1636772, Temple University Data Science Targeted Funding Program, NSF grant CNS-1625061, Pennsylvania Department of Health CURE grant and ONR/ONR Global (grant No. N62909-16-1-2222).

References

[1] Qin, T., Liu, T.Y., Zhang, X.D., Wang, D.S., Li, H.: Global ranking using continuous conditional random fields. In: Advances in neural information processing systems. pp. 1281-1288 (2009)

[2] Radosavljevic, V., Vucetic, S., Obradovic, Z.: Continuous Conditional Random Fields for Regression in Remote Sensing. In: ECAI (2010)

[3] Zaas, A.K., Chen, M., Varkey, J., Veldman, et al.: Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell host & microbe 6(3), 207-217 (2009)

Corresponding e-mail: martin.pavlovski@temple.edu

ECML/PKDD 2017, THE EUROPEAN CONFERENCE ON MACHINE LEARNING and PRINCIPLES & PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES, SKOPJE, MACEDONIA, SEPTEMBER 2017