Structured regression on partially observed evolving graphs with uncertainty propagation

Jelena Stojanovic
Temple University
Philadelphia, PA 19122
jelena.stojanovic@temple.edu

Djordje Gligorijevic
Temple University
Philadelphia, PA 19122
gligorijevic@temple.edu

Milos Jovanovic
University of Belgrade
11000 Belgrade, Serbia
milos.jovanovic@fon.bg.ac.rs

Zoran Obradovic
Temple University
Philadelphia, PA 19122
zoran.obradovic@temple.edu

Abstract

Conditional probabilistic graphical models provide a powerful framework for structured regression in spatio-temporal datasets with complex correlation patterns. However, in real-life applications a large fraction of observations is often missing, which can severely limit the representational power of these models. We have proposed a Marginalized Gaussian Conditional Random Fields (m-GCRF) structured regression model for dealing with missing labels in partially observed temporal attributed graphs. This method is aimed at learning with both labeled and unlabeled parts and effectively predicting future values in a graph. The method is even capable of learning from nodes for which the response variable is never observed in history, which poses problems for many state-of-the-art models that can handle missing data. The proposed model is characterized for various missingness mechanisms on 500 synthetic graphs. The benefits of the new method are also demonstrated on a challenging application for predicting precipitation based on partial observations of climate variables in a temporal graph that spans the entire continental US. We also show that the method can be useful for optimizing the costs of data collection in climate applications via active reduction of the number of weather stations to consider. In experiments on these real-world and synthetic datasets we show that the proposed model is consistently more accurate than alternative semi-supervised structured models, as well as models that either use imputation to deal with missing values or simply ignore them altogether. Additionally, structured regression models are applicable to high impact applications such as healthcare and medicine. However, having good prediction accuracy alone is often not enough. These kinds of applications require a decision making process which uses uncertainty estimation as input whenever possible. Quality of uncertainty estimation is subject to over or under confident prediction, which is not addressed in many models. We have proposed extensions for the GCRF model that have been applied to the temporal disease graph built from the State Inpatient Database (SID) of California, acquired from the HCUP. Our experiments demonstrated benefits of using graph information in modeling temporal disease properties as well as improvements in uncertainty estimation provided by given extensions of the Gaussian Conditional Random Fields method. For the applications of predicting climate and healthcare trends far in the future we have developed an uncertainty propagation framework for Gaussian structured models that has shown great improvements over non-uncertainty propagation models as well as different state-of-the art uncertainty propagation regression models.

Learning and inference with partially observed data is a challenge experienced in many real-world domains and one of the main problems with partial observation, especially of temporal data, is the inability to observe all graph attributes throughout the entire history of the graph. In this phase we are dealing with the problem of being unable to observe the target variable (that we want to model) in all nodes. In particular, we address the problem of structured regression in a temporal graph (prediction of continuous node states in time step $t + 1$), where the dependent variable (label) is missing in a large fraction (up to 80%) of training data (time points $1, 2, ..., t - 1, t$). This constitutes a semi-supervised learning (parameter estimation) problem, which is distinct from approaches that try to infer the labels of the unlabeled nodes of a graph. In this period an even more challenging problem is considered, where labels at some nodes are missing at all time steps. The nodes in a graphical model are not independent, and so ignoring training data with missing labels might disregard too much information. If nodes with missing labels are ignored, the entire graph structure could be lost and modeling would be limited to unstructured regression or time-series prediction on individual nodes. Utilizing the graph structure may therefore make better use of unlabeled data, especially when lots of nodes have missing labels.

We have tested the proposed method (and selected benchmarks) against different missing mechanisms that can naturally occur, including labels missing completely at random, and missing labels influenced by node attributes, by network connections, or by previous missing values. We have evaluated the described methods using:

- synthetic data (about 500 spatio-temporal graphs with up to 15,000 nodes in 5 time steps)
- real-world climate application for precipitation prediction, where the missing labels are present in the observed graph history we use for training the model.

The results are reported in terms of mean and standard deviation of R^2 as the accuracy measure for 0 to 80% of missingness in data for the proposed m-GCRF model and previously mentioned benchmark models. Our experiments provide evidence that, under these various missingness mechanisms, the proposed approach is more effective then other alternatives.

The GCRF model we are using for structured regression, intrinsically possesses uncertainty estimation. However, the uncertainty estimation is highly biased towards the unstructured predictors the model was trained on, as GCRF does not depend on input variables directly. The parameters of GCRF represent the degree of belief in each unstructured predictor and graph structure. However, the model does not capture the uncertainty of the unstructured predictors. Therefore, we extended it to take into consideration the uncertainty of the unstructured predictors and distribution of input features. The resulting model is called uGCRF. The parameters α (degrees of belief towards unstructured predictors that the model is using) are modeled as both parameterized input variables $\alpha(\theta_k, x)$, where θ_k are the parameters, and x are input variables of function α, and a non parametrized function of the uncertainty of each unstructured prediction. The principal assumption is that the chosen unstructured predictors can output uncertainty for their predictions. Both proposed models have shown significant improvement in uncertainty estimation compared to the both original model and other non structured time series models on the disease group phenotype graph built from California State Inpatient database obtained from HCUP.

1 Acknowledgments

This research was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, National Science Foundation through major research instrumentation grant number CNS-09-58854

References

[1] Gligorijevic, Dj. & Stojanovic, J. & Obradovic, Z. (2015) Improving Confidence while Predicting Trends in Temporal Disease Networks, *4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining, Vancouver, Canada, April 30 - May 02, 2015*

[2] Stojanovic, J. & Jovanovic, M. & Gligorijevic, Dj. & Obradovic, Z. (2015) Semi-supervised learning for structured regression on partially observed attributed graphs *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015) Vancouver, Canada, April 30 - May 02, 2015*