Graph Regularized Meta-path Based Transductive Regression in Heterogeneous Information Network

Mengting Wan, Yunbo Ouyang, Lance Kaplan, Jiawei Han

University of Illinois at Urbana-Champaign, U.S. Army Research Laboratory

mwan5@illinois.edu, youyang4@illinois.edu, lance.m.kaplan.civ@mail.mil, hanj@illinois.edu

Abstract

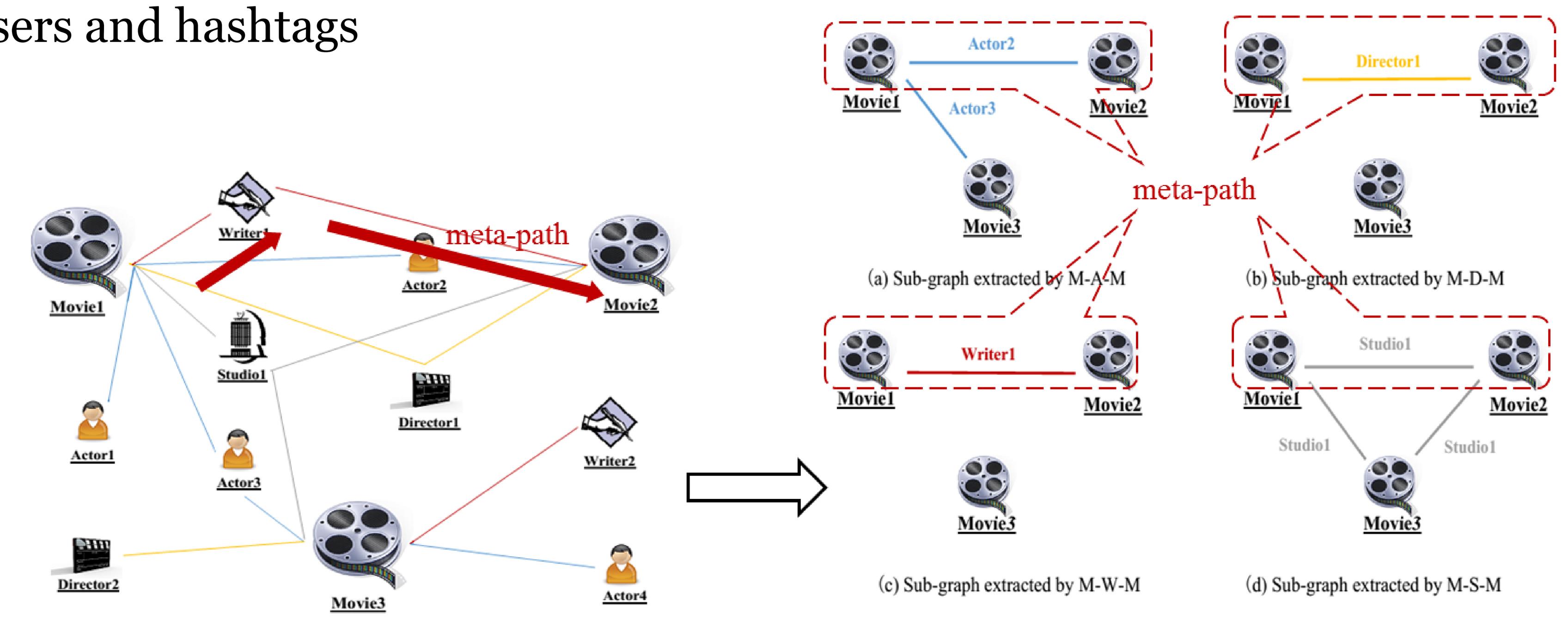
A number of real-world networks are heterogeneous information networks (HIN), which are composed of different types of nodes and links. **Numerical prediction in HIN** is a challenging but significant area because network based information for unlabeled objects is usually limited to make precise estimations. In this paper, we consider a graph regularized meta-path based transductive regression model *(Grempt)*, which combines the principal philosophies of typical graph-based transductive classification methods [1,2] and transductive regression models designed for homogeneous networks [3]. The computation of our method is time and space efficient and the precision of our model can be verified by numerical experiments.

Introduction

Heterogeneous Information Network (HIN) is a kind of information network where objects and links have different types. Numerical Prediction in HIN is aimed to predict numerical attributes based on the HIN structure.

Examples of Numerical Prediction in HIN:

- Predict box-office and expected rating score of an upcoming movie based on an IMDb network
- Predict the total number of citations of an author based on the DBLP plus citation network
- Predict the number of retweets based on twitter network composed of tweets, users and hashtags



Model

We proposed a graph regularized meta-path based transductive regression model *(Grempt)*.

Optimization Framework:

Included in most classification frameworks

objective function = *graph regularization + loss on labeled objects*

+ *loss on unlabeled objects (pseudo-labels involved)*

Designed for numeric prediction

$$\min_{\mathbf{f},\mathbf{w}} \ \mathbf{J}(\mathbf{w};\mathbf{f}) = \boldsymbol{\Omega}(\mathbf{w};\mathbf{f}) + \alpha_1 \mathbf{C}_1(\mathbf{f}_L;\mathbf{y}_L) + \alpha_2 \mathbf{C}_2(\mathbf{f}_U;\tilde{\mathbf{y}}_U)$$

$$= \boxed{\sum_{k=1}^{K} w_k} \left[\sum_{u,v=1,u\neq v}^{m+n} R_{uv}^{(k)} \left(\frac{f_u}{\sqrt{D_u^{(k)}}} - \frac{f_v}{\sqrt{D_v^{(k)}}} \right)^2 \right]$$

weights of different types of meta-paths

$$+ \alpha_1 \sum_{u=1}^{n} (f_u - y_u)^2 + \alpha_2 \sum_{v=1}^{m} \frac{(f_{n+v} - \tilde{y}_{n+v})^2}{\sigma^2_{\tilde{y}_{n+v}}}$$

subject to

$$\sum_{k=1}^{K} \exp(-w_k) = 1. \tag{4.5}$$

Three principles:

- predictions of the target variable of two linked objects are likely to be similar — *graph regularization*
- predictions of the target variable of labeled objects should be similar to their labels — *loss on labeled objects*
- predictions of the target variable of unlabeled objects should be similar to their local estimated labels (pseudo-labels) — *loss on unlabeled objects*

Algorithm:

- Determine pseudo-labels of unlabeled objects and their associated variance using local information

$$\tilde{y}_{n+v} = \sum_{u\in\mathcal{N}_q(x_{n+v})} p_{n+v,u} y_u = \frac{\sum_{u\in\mathcal{N}_q(x_{n+v})} R_{n+v,u} y_u}{\sum_{u\in\mathcal{N}_q(x_{n+v})} R_{n+v,u}}. \tag{4.6}$$

- Initialize numeric predictions $\mathbf{f}$ and weights of meta-path $\mathbf{w}$
- Iteratively update $\mathbf{f}$ and $\mathbf{w}$ until converge
  - Suppose $\mathbf{f}$ is fixed, we can obtain a closed form solution for the weights of meta-path $\mathbf{w}$
  - Suppose $\mathbf{w}$ is fixed, we can obtain the solution of $\mathbf{f}$ by solving an linear equation system or by a iterative method

Experiment

Datasets:

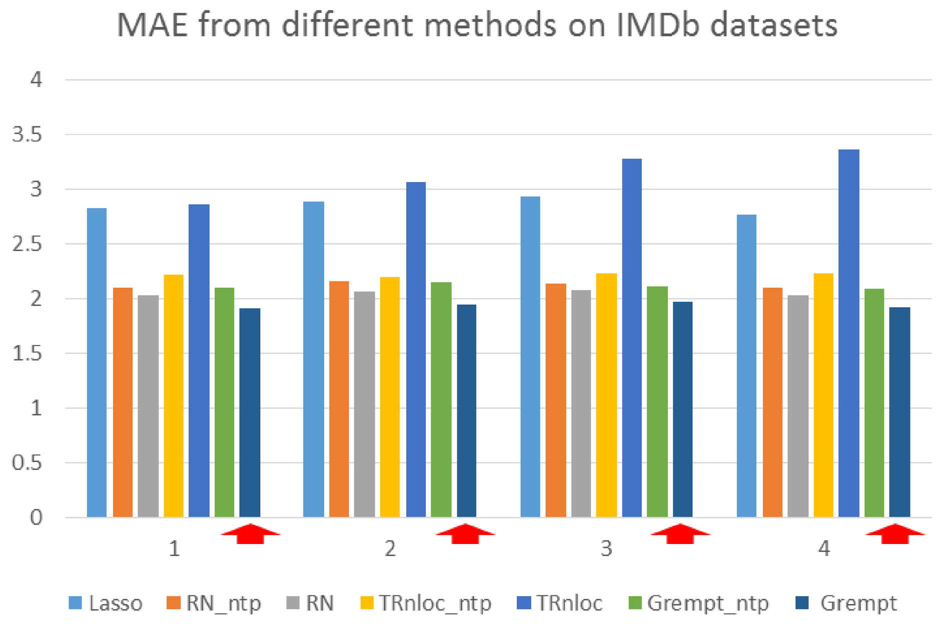
- IMDb: predict box-office sales of movies
- DBLP: predict total number of citations of authors

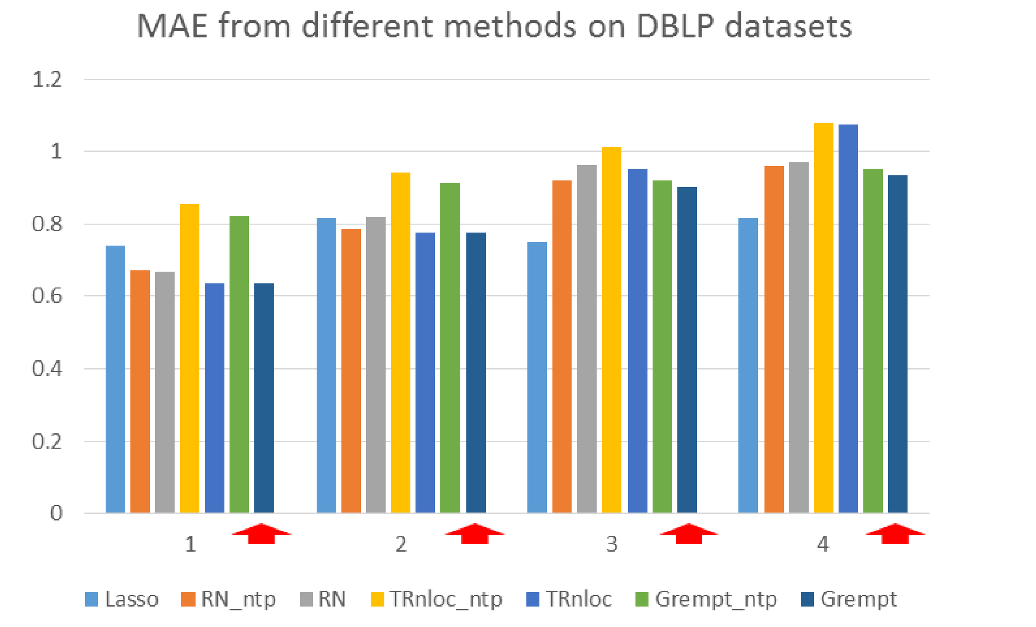
IMDb	Number of labeled objects	Number of unlabeled objects	Percentage of labeled objects
dataset1	3067 (2000–2012)	233 (2013–2013)	92.94%
dataset2	2820 (2000–2011)	480 (2012–2013)	85.45%
dataset3	2578 (2000–2010)	722 (2011–2013)	78.12%
dataset4	2345 (2000–2009)	955 (2010–2013)	71.06%
DBLP	**Number of labeled objects**	**Number of unlabeled objects**	**Percentage of labeled objects**
dataset1	3017	315	90.55%
dataset2	1666	1666	50.00%
dataset3	334	2998	10.02%
dataset4	167	3165	5.01%

Table 1: Summary of IMDb datasets (numbers in parentheses indicate released year) and DBLP datasets.

Methods for Comparison:

- LASSO [4]
- Relational neighbor estimation with/without type information — RN_ntp/RN_tp [5]
- Transductive regression without penalty of local estimates with/without type information — TRnloc_ntp/TRnloc [1, 2]
- Graph regularized meta-path based transductive regression with/without type information — *Grempt_ntp/Grempt* (Our method)





$$MAE = \frac{1}{m}\sum_{v=1}^{m} |f_{n+v} - y_{n+v}|.$$

Results:

- Our *Grempt* model has the best performance based on Mean Absolute Error.
- Those important meta-paths will increase as the number of labeled objects decreases while contributions of relatively unimportant meta-paths will decrease.

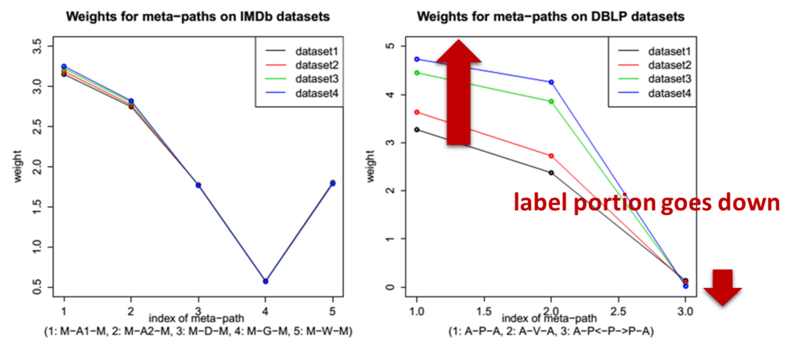


Figure 3: Weights for meta-paths of IMDb datasets and DBLP datasets from *Grempt* Model.

Reference

[1] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, "Graph regularized transductive classification on heterogeneous information networks," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 570–586.

[2] M. Ji, J. Han, and M. Danilevsky, "Ranking-based classification of heterogeneous information networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1298–1306.

[3] C. Cortes, M. Mohri, and M. Mohri, "On transductive regression," in *NIPS*, 2006, pp. 305–312.

[4] R. Tibshirani, "Regression shrinkage and selection viathe Lasso," *Journal of the Royal Statistical Society*. Series B (Methodological), pp. 267–288, 1996.

[5] S. A. Macskassy and F. Provost, "A simple relational classifier," DTIC Document, Tech. Rep., 2003.

ILLINOIS ARL