Node Classification in Dynamic Social Networks

Yulong Pei Y.PEI.1@TUE.NL
Jianpeng Zhang J.ZHANG.4@TUE.NL
George Fletcher G.H.L.FLETCHER@TUE.NL
Mykola Pechenizkiy M.PECHENIZKIY@TUE.NL
Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands

Keywords: Node Classification, dynamic Factor Graph Model, Dynamic Social Networks

Abstract

In this paper, we propose the *dynamic Factor Graph Model* (dFGM), which is an extension of Factor Graph Models, for the problem of node classification in dynamic social networks. *dFGM* can capture not only node attributes and correlations but also the temporal information for node classification. We conduct experiments on a real-world data set and the experimental results demonstrate the effectiveness of our proposed method.

1. Introduction

Recently the problem of node classification in networks has attracted extensive attention and the independent and identically distributed (*i.i.d.*) assumption does not hold for it. Methods should take the structure dependency into account. In practice, many real-world networks are dynamic and nodes/edges in the networks may change during time. In such dynamic scenario, temporal information can also play an important role in classifying nodes. Therefore, in this paper we propose the *dynamic Factor Graph Model* (dFGM) for node classification in dynamic social networks. In detail, the dynamic graph data is organized in the format of a series of graph snapshots. To model the graph snapshots, three types of factors, i.e., node factor, correlation factor and dynamic factor, are designed in the *dFGM* based on node features, node correlations and temporal correlations, respectively. Node factor and correlation factor can capture the global and local properties of the graph structures while the dynamic factor can make use of the temporal information.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

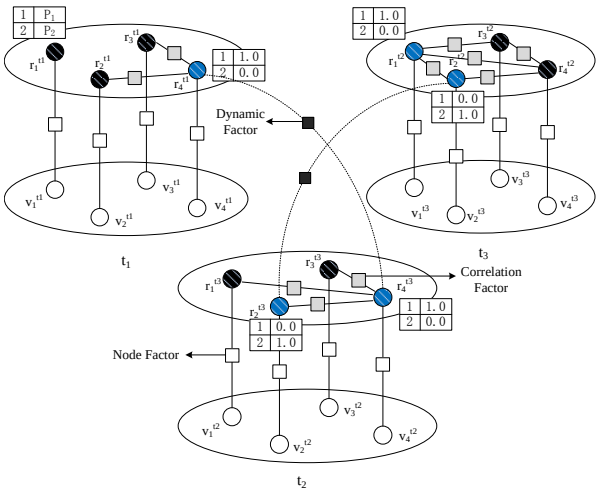


Figure 1. A four-node example in three snapshots of the *dFGM*. The white circles in the lower layers denote the nodes, the colored circles in the upper layers denote the labels and the squares are the factors.

2. Dynamic Factor Graph Model

Factor graph models (FGM) (Koller & Friedman, 2009) have been widely used in modeling dependency between entities that can be organized in a graph structure and are flexible in integrating with task-specific functions. In this section, we present the proposed *dynamic Factor Graph Model* (dFGM), which is an extension of the traditional FGM, for node classification in dynamic social networks.

Intuitively the class/label of a node in a dynamic social network will be determined by three factors: (1) *node attributes*, i.e., the node's global and local characteristics at current time step. This factor corresponds to the node features extracted from the network or nodes' intrinsic characteristics; (2) *node correlations*, i.e., the relationships or interactions between nodes in a snap-

Table 1. Experimental results.

Methods		*accuracy*	*error*
Feature-based classification	LR	0.3157±0.0031	———
	SVM	0.4983±0.0029	1.2651±0.0027
Link-based classification	MV+dynamic	0.5049+0.0011	0.9068+0.0009
	CC	0.7935+0.0033	0.8642±0.0101
Factor Graph Models	FGM_feat	0.2684±0.0024	1.4917±0.0003
	FGM_corr	0.8360±0.0328	0.7865±0.0062
	dFGM	**0.8410±0.0058**	**0.7360±0.0073**

shot. This factor corresponds to the social relations of nodes in the networks; (3) *previous labels*, we assume that the label of a node will not change abruptly. This factor can capture the dynamic information in the network.

Based on these intuitions, we propose the *dFGM* which consists of three factors, named node factor, correlation factor and dynamic factor, and they correspond to the node attributes, node correlations and previous labels, respective. In detail,

- Node factor $f(r_i, \boldsymbol{x}_i)$. This factor represents the posterior probability of the label r_i give the feature $\mathbf{x}_i$ of node v_i.
- Correlation factor $c(r_i, N(r_i))$. This factor reflects the correlation between nodes, where $N(r_i)$ is the set of correlated labels to r_i. There are multiple ways to define the set of correlated labels and in this study, $N(r_i)$ denotes the labels of neighbors of node v_i.
- Dynamic factor $d(r_i^t, r_i^{t-1})$. This factor denotes the correlation between the labels of one node in two consecutive snapshots.

An example of the *dFGM* with three factors is shown in Fig 1. Based on all the factors introduced above, the joint distribution of labels R given the graph $\mathcal{G}$ can be defined as

$$P(R|\mathcal{G}) = \prod_t \prod_i f(r_i, \boldsymbol{x}_i) c(r_i, N(r_i)) d(r_i^t, r_i^{t-1}) \quad (1)$$

To simplify the model learning, we choose the exponential-linear function to instantiate the factors. In detail, the node factor is defined as

$$f(r_i, \boldsymbol{x}_i) = \frac{1}{Z_1} \exp\{\alpha^T \boldsymbol{\phi}(r_i, \boldsymbol{x}_i)\} \quad (2)$$

where Z_1 is the normalizing factor, α is the weighting vector, and $\boldsymbol{\phi}$ is a vector of feature function. Similarly, the edge factor is defined as

$$c(r_i, N(r_i)) = \frac{1}{Z_2} \exp\{\sum_{r_j \in N(r_i)} \beta^T \mathbb{I}_{corr}(r_i, r_j)\} \quad (3)$$

where Z_2 is the normalizing factor, β is the weighting vector, and $\mathbb{I}_{corr}$ is the indicator function for node correlations and defined as

$$\mathbb{I}_{corr}(r_i, r_j) = \begin{cases} 0, & \text{if } e_{ij} \notin E \\ 1, & \text{if } e_{ij} \in E \end{cases} \quad (4)$$

Then temporal factor is defined in the same way

$$d(r_i^t, r_i^{t-1}) = \frac{1}{Z_3} \exp\{\gamma^T \mathbb{I}_{dyn}(r_i^t, r_i^{t-1})\} \quad (5)$$

where Z_2 is the normalizing factor and γ is the weighting vector. $\mathbb{I}_{dyn}(r_i^t, r_i^{t-1})$ is the indicator function for the dynamic information and defined as

$$\mathbb{I}_{dyn}(r_i^t, r_i^{t-1}) = \begin{cases} 0, & \text{if } r_i^t \neq r_i^{t-1} \\ 1, & \text{if } r_i^t = r_i^{t-1} \end{cases} \quad (6)$$

3. Experiments

We conduct the experiments to validate the performance of *dFGM* on the DBLP data set. We extract data from six research areas from 2001 to 2010 and use these areas as classification labels. We use two evaluation metrics in experiments, i.e., *accuracy* and *error* in probability. The *accuracy* is $accuracy = n/N$ where n is the number of instances correctly classified, and N is the total number of instances in the test data. The *error* in probability is defined as $error = \sum_{i=1}^{N} \sum_{j=1}^{c} |\hat{p}_i^j - p_i^j|/N$ where N is the number of instances in the test data and c is the number of labels. $\hat{p}_i^j$ and p_i^j are the predicted probability and ground probability of label j for user i, respectively.

To validate the effectiveness of our proposed *dFGM*, we compare three types of methods: feature-based classification methods (LR and SVM), link-based classification methods (MV+dynamic and CC) and factor graph model based methods (FGM_feat, FGM_corr and *dFGM*). The experimental results are shown in Table 1. Note that the features used here are extracted using DeepWalk (Perozzi et al., 2014).

References

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). Deepwalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701–710).