Evolution of Node Behavior in Link Prediction

Baojun Qiu[1], **Qi He**[2], **John Yen**[3]
Department of Computer Science and Engineering[1]
College of Information Science and Technology[3]
Pennsylvania State University, University Park, 16801
IBM Almaden Research Center, San Jose, CA 951203[2]
bqiu@cse.psu.edu, heq@us.ibm.com, jyen@ist.psu.edu

Introduction

In the study of social network evolution, one of the central tasks is *link prediction* to infer new links between nodes. Link prediction has many applications, including recommending new items in online networks (e.g., products in eBay and Amazon, and friends in Facebook), monitoring and preventing criminal activities in a criminal network, predicting the next webpage users will visit, and complementing missing links in automatic web data crawlers.

Link prediction problems are often converted to supervised learning tasks. Node pairs without links in a snapshot of a social network at time t are sampled as training samples. If a node pair forms a new link in the snapshot at the next time step, then it is a positive sample. Otherwise, it is a non-positive sample. The features of a sample are calculated on the snapshot where the node pair is sampled. In general, the features include topological features such as degree and distance (Liben-Nowell and Kleinberg 2007). Classifiers (or rankers) are learned from the training samples.

Evolution of node behavior is observed in many social networks and has been shown to be useful in modeling social networks (Qiu et al. 2010). For example, in a scientific collaboration network, a student researcher tends to work with senior researchers; as she establishes a career, her collaboration preference becomes more diverse. This suggests the behavior of nodes may have phase change and temporal trends. Intuitively, taking behavior evolution of nodes into consideration may improve link prediction. However, little work has been done, probably due to the difficulty in describing and characterizing node behavior and its evolution.

In this paper, we use time series to describe node behavior, calculate temporal features from the time series to characterize behavior evolution, and use the temporal features to improve link prediction.

Methodology

In traditional link prediction approaches, the features of a sample (a node pair) are calculated in the snapshot where the node pair is sampled. In our approach, to detect potential evolution, we instead calculate features of a sample on all previous snapshots. Therefore, each sample is a vector in traditional approaches, and each element is a numerical value (only consider numerical features), while in our approach, each sample is still a vector, but each element is a $time\ series$.

Classifying vectors of time series is a hard problem. We decided to extract temporal features from time series. On one hand, temporal features characterize the evolution and temporal trends. On the other hand, samples (vectors of time series) are converted into traditional samples (vector of temporal features), and existing learning methods can be used.

We consider the following four types of temporal features to characterize time series.

Simple Statistics. This type of feature includes simple first-order temporal features such as $recency$ (Potgieter et al. 2007) and $activeness$ (Huang and Lin 2009). $recency$ measures the length of time elapsed since a node made its last connection. A large $recency$ indicates that a node has been inactive for a long time, and likely to be inactive in future. $activeness$ measures the number of connections made by a node in the latest time step. A large $activeness$ indicates that a node is very active in the last time step and is likely to be active in the future.

Local Pattern. This type of feature calculates whether a time series has a particular local pattern, e.g., a time series first decreases for 5 time steps and then increases for 5 steps, and thus has a V-shape. These features are useful to detect phase changes of node behavior. To calculate this type of feature, we extend the approach proposed by Kadous (Kadous 1999).

Predicition. This type of feature captures global trends. The global trend of a time series is usually complex and cannot be simply portrayed by a single pattern. We use the 1-step ahead prediction of a time series to partially describe the global trend. The intuition is: if two (normalized) time series have the same global temporal trend, they have the same 1-step ahead prediction value.

Interplay. The problem of link prediction is to predict link formation between two nodes. We thus especially define *Interplay*, which computes the joint likelihood of two nodes to be connected in the future based on to which degree the two nodes match each other's preference. Suppose we focus on their preference on $degree$, and A and B are two nodes. In the last three time steps, the average degrees of nodes connected by A at each time step were 10, 20, 30, then we

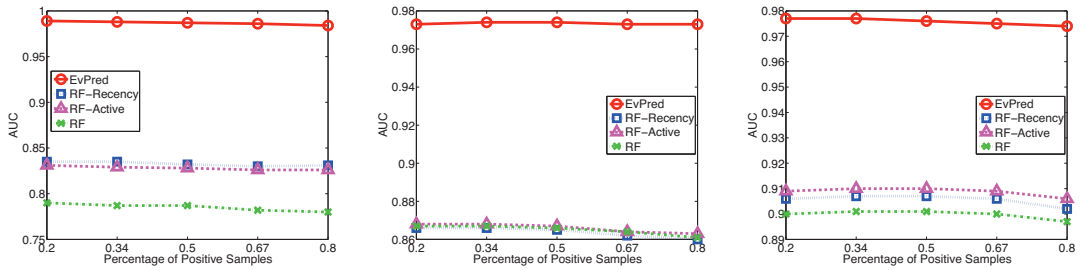

Figure 1: Performance(AUC) of different models on different data: NanoSCI (Left), Phone (Middle), Facebook (Right)

may predict the degree of nodes that A will connect to in the coming time step is a distribution (N_A) with mean around 40. If we predict that the degree of B (D_B) to be around 40, then B has a high likelihood to be chosen by A as a new neighbor. We therefore define the joint likelihood focusing on $degree$ as $L_{degree}(A, B) = f(D_B; N_A)f(D_A; N_B)$, where $f(x; dist)$ indicates the probability density of x in distibution $dist$, N_B is defined as the predicted distribution of the degree of B's new neighbors, and D_A is the predicted degree of A in the coming time step.

After the features are calculated for samples, we train learners for link prediction.

Evaluation

We tested three real social networks. **NanoSCI** is a scientific collaboration network in the nanotechnology research community from 1980 to 2006 (292,323 nodes, each year is a time period). **Phone** is a cell phone communication network in a European country from 09/2007 to 03/2008 (24,986 nodes, each week is a time period). **Facebook** is the wall-to-wall post relationship on Facebook.com from 09/2006 to 01/2009 (66,842 nodes, each month is a time period).

For each network, we extract training samples from the first two thirds snapshots, and testing samples from the rest of the snapshots. Note that social networks are sparse such that positive links only occupy a small portion of all pair of nodes. To make the training and testing data balanced, randomly sampling of negative samples is used in literature so that the number of negative samples is comparative to that of positive samples. This kind of random sampling is controversial. First, the training and testing data are not representative for the underlying social networks. Second, the testing accuracy is higher than expected. For example, the distances in negative samples (node pairs) are usually significantly larger than the distances in positive samples. In our experiments, we found that setting a threshold on distance achieve accuracies around 70-80% on the three datasets. However, in this paper, we still use randomly sampling for negative samples because our focus is the comparison of our approach and existing approaches.

We consider three baselines. RF is an evolution-agnostic random forest-based approach proposed by Lichtenwalter $et\ al.$ (Lichtenwalter, Lussier, and Chawla 2010). It only includes the static topological features. RF-Recency is an modified version of the approach proposed by Potgieter $et\ al.$ (Potgieter et al. 2007). It uses $recency$ as well as the static topological features. RF-Active uses $activeness$ as well as the static topological features. Both RF-Recency and RF-Active are also random forest-based models. Our model, Ev-Pred, is also random forest-based and include all temporal features and static topological features.

Figure 1 shows results measured in AUC on three datasets (The AUC is high for all models particially because of randomly sampling of negative samples, and AUC of RF agrees with that reported in Lichtenwalter's paper). For each of them, we control the percentage of positive sample in both training and testing as 1/5, 1/3, 1/2, 2/3 and 4/5 respectively. The results suggest that our approach consistently and significantly outperforms the others on all datasets at all ratio of positive samples. It suggests that temporal features describing the evolution of node behavior are extremely useful to the problem of new link prediction.

Conclusion

We calculated temporal features to characterize the evolution of node behavior, and our experimental results suggest that including these temporal features significantly improve link prediction performance.

Acknowledgments

This work was supported by a grant from the Defense Threat Reduction Agency (HDTRA1-09-1-0054). We thank Prof. Frank E. Ritter and the anonymous reviews for helpful comments.

References

Huang, Z., and Lin, D. 2009. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*.

Kadous, M. W. 1999. Learning comprehensible descriptions of multivariate time series. *ICML'1999*.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*.

Lichtenwalter, R.; Lussier, J.; and Chawla, N. 2010. New perspectives and methods in link prediction. *SIGKDD'2010*.

Potgieter, A.; April, K.; Cooke, R.; and Osunmakinde, I. 2007. Temporality in link prediction: Understanding social complexity. *Sprouts: Working Papers on Info. Sys.*

Qiu, B.; Ivanova, K.; Yen, J.; and Liu, P. 2010. Behavior evolution and event-driven growth dynamics in social networks. *SocialCom'2010*.