The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Social Influence Prediction by a Community-based Convolutional Neural Network

Shao Hsuan Tai[*1] Hao-Shang Ma[*2] Jen-Wei Huang[*3]

[*1*2*3]Institute of Computer and Communication Engineer,
Department of Electrical Engineering,
National Cheng Kung University, Tainan, Taiwan

Learning social influence between users on social networks has been extensively studied in a decade. Many models were proposed to model the microscopic diffusion process or to directly predict the final diffusion results. However, most of them need expensive Monte Carlo simulations to estimate diffusion results and some of them just predict the size of the spread via regression techniques, where people who will adopt the information becomes unknown. In this work, we regard the prediction of final influence diffusion results in a social network as a classification problem to avoid expensive simulations with knowing the final adopters. We first address the problem on a deep neural network and utilize the diffusion traces to train the network. Furthermore, we propose a community-based convolutional neural network to capture the information of local structure with the aforementioned network. The proposed model is referred to as the Social Influence Learning on Community-based Convolutional Neural Network, SIL-CCNN. In the experiment, SIL-CCNN shows the promising results in both synthetic and real-world datasets.

1. Introduction

Nowadays people tend to share their life, emotions, and opinions to others on social network websites. Two representative diffusion models, Independent Cascade, IC, model and Linear Threshold, LT, model, were reformulated by Kempe [3]. However, there are several limitations of predicting the information diffusion using diffusion models. The influence probabilities between users and the active threshold of a user should be measured or learned from many personal features such as users' preferences and different relationships. In real social networks, the features are not easy to extract since the data sometimes is not complete. In addition, to get the results of IC/LT model, we need to conduct a huge number of simulations.

Actually, the prediction of the information diffusion process and the final diffusion results models can be regarded as a classification problem. Given the information sources and the network structure, the individuals are classified into active or inactive classes in the final diffusion result. The active class is corresponding to the individuals being influenced successfully in the information diffusion process. Some related works aim to predict the size of information spreads as a classification or regression problem [1, 7]. Different with diffusion models, these methods usually do not learn the individuals who are actually influenced by the information. We would like to know exactly who are influenced and who are not.

To overcome above limitations and solve the classification problem, we propose an influence prediction model based on community-based convolutional neural network. First, we consider the influence propagation process in the past to learn the influence between individuals. Then, most of the diffusion models consider the network structure, i.e., the relations between individuals, as their features. However, the information of the whole network structure may not be useful for the classification problem whereas the local network information of a single individual should be helpful. We try to embed the local structure of an individual into our model to learn the local relations. The community structure in social networks represents a cluster of individuals sharing connections that are stronger than those with individuals outside the community. The information diffusion in a community should be faster than outside the community. Therefor, we use the convolutional neural networks to mine the local relations within the community structure. The idea is that convolutional neural networks are good to extract the effects of a small group of individuals around an individual by extracting valuable relations from the structure. Finally, the influence traces and the community structure information are combined into our model as training features of the deep neural network. The proposed scheme is referred to as the Social Influence Learning on Community-based Convolutional Neural Network, SIL-CCNN.

The remainder of the paper is organized as follows. Works related to this work are outlined in Section 2. Section 3 details the proposed methodology. Experiment results and conclusions are presented in Section 4 and Section 5.

Contact: Jen-Wei Huang, Institute of Computer and Communication Engineer,
Department of Electrical Engineering,
National Cheng Kung University, Tainan, Taiwan.
Email: jwhuang@mail.ncku.edu.tw
Tel:(+886)-6-2757575#62347

2. Previous Work

In this section, we will briefly introduce other related works on diffusion models and the prediction of information spread size.

2.1 Diffusion Models

Diffusion models can be used to identify social influence by tracking information propagating through a social network. Nowadays, some diffusion models adopt the learning strategy to predict the activation since the technique of learning methodology has matured in these few years. Saito [6] first proposed a learning method for IC model. Wang [8] proposed a feature-enhanced approach, which considers not only temporal data in cascades but also additional features. Chou [2] proposed Multiple Factor-Aware Diffusion model, MFAD, that can consider many kinds of factors together. MFAD model adopts positive and unlabeled learning to train the classifiers for each individual.

2.2 Prediction of Information Spread Size

Different from diffusion models, the cascade prediction only focuses on predicting whether the information will become popular and widely spread. The problem is usually formulated as a classification or a regression problem to understand the size of potential influence of information.

Among the classification solutions of predicting information cascade, Cheng [1] proposed to use temporal and structural features for predicting the relative growth of a cascade size. As for the regression solutions, Tsur [7] proposed a content-based prediction model to include locations, orthography, number of words, lexicality, ease of cognitive process and emotional effect on various cognitive dimensions.

Our work aims to learn the hidden influence propagation from the data instances and the network structure directly without the huge number of diffusion simulations.

3. Methodology

In this section, we first propose an ordinary deep neural network model, SIL-DNN, for identifying traces of social influence. Second, we adopt a convolutional neural network to extract the local structure information of a network from the communities. Then, we propose Social Influence Learning on Community-based Convolutional Neural Network, abbreviated as SIL-CCNN, which combines the SIL-DNN and a community-based convolutional neural network to predict the final influential results.

Given a social network $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, the node set $\mathbb{V}$ corresponds to the individuals, and $\mathbb{E}$ is the edge set indicating the relationships between individuals. Each influence trace (u, v, x) indicates that the nodes v adopt the information x and is influenced by source nodes u, where u and v are sets of nodes in $\mathbb{V}$. The architecture of SIL-DNN is presented in Fig. 1. In SIL-DNN, the number of neurons in the input layer is the same as the number of individuals in the social network $|V|$. The same setting is used in the output layer. For every trace, we put the vector of information sources as the input data of SIL-DNN and the output are the vector of influenced nodes. The input vector and output vector could be set as follows,

$$\begin{cases} y_i = 1, \text{if } i \in u \text{ for } (u, v, x) \\ y_i = 0, \qquad \text{otherwise,} \end{cases} \tag{1}$$

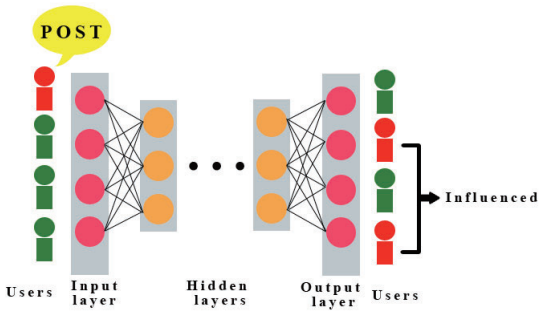


Figure 1: Social Influence Learning on Deep Neural Network Architecture

where y_i represents an individual i at the input layer. $i \in u$ for (u, v, x) indicates all traces that transmit the information x from user i. For example, an individual i provides an information x during the observation. The neuron of input layer $y_i^0 = 1$ for the information x. On the other hand, in the output layer, the $y_j = 1$ represents that node j is influenced by i and is classified as the active class. Otherwise, node j is classified into the inactive class.

3.1 Social Influence Learning on Community-based Convolutional Neural Network

In order to join the community structure to help us to predict the information diffusion results, we propose Social Influence Learning on Community-based Convolutional Neural Network, SIL-CCNN. The architecture of SIL-CCNN is presented in Fig. 2. First, we need to extract the community information in the network and form the relation matrix of each community. A list of relation matrix RM for communities in $COMM$ can be formulated. Then, we design a community-based convolutional neural network to deal with community-related information by extracting features through a convolutional layer and a pooling layer. For the input of the convolution layer, we extract the relation matrices of communities to represent the local network information of the individual. The relation matrix RM of a community is defined as follows,

$$RM = [rm_{ij}]_{d \times d}, rm_{ij} = \begin{cases} w_{ij}, (v_i, v_j) \epsilon E \\ 0, otherwise, \end{cases} \tag{2}$$

where d is the number of nodes in the community. The element rm_{ij} in the relation matrix represents the weight on the edge between node v_i and node v_j in a community. In addition, in order to account for differences in the size of communities, the community information matrices are normalized to the size of the largest community and have zero-padding.

However, if we randomly assign the order of nodes in relation matrices and put the matrices into the convolutional neural network, the small extracted region in the convolution layer would be meaningless. Therefore, we design an arrangement strategy to determine the relations of nearby

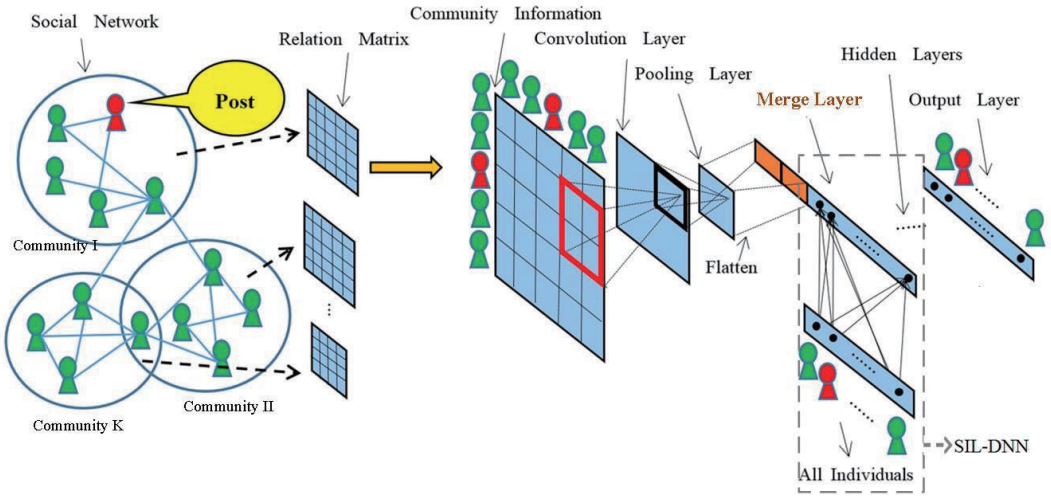


Figure 2: Social Influence Learning on Community-based Convolutional Neural Network architecture

elements. The order of individuals in each relation matrix is determined by the ranking of the number of degrees. The individual having the largest degree will be arranged to the leftmost of x-axis and the topmost of y-axis. Therefore, each block in different relation matrices shares the same weight matrix in the convolutional neural network.

In the pooling layer, we use max-pooling as our pooling function. Max-pooling is particularly well suited to the separation of features that are sparse. After the pooling layer, SIL-CCNN constructs a merge layer to combine the output of the pooling layer and the input vector of SIL-DNN. Then, several hidden layers are trained in SIL-CCNN before the output layer. Finally, a few hidden layers and one full-connected output layer are connected after the merge layer.

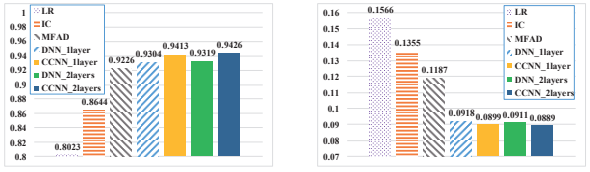
As for the detailed setting of neural network, the kernel size of the convolutional layer is defined to 3×3. In addition, we define a sigmoid function in the output layer and a cross-entropy objective function. In the training step, our goal is to minimize the following equation:

$$Q\left(W^t, W^r\right) = -\sum_{i}^{I} t_i \log y_i, \tag{3}$$

where I is the number of inputs, W^t is the weight matrices of traces, and W^r represents the weight matrices of the structure relation in SIL-CCNN. The weight matrix contains the relations between the individuals and the local structure information. The equation above is shown for one individual i at output layer k ranging over all target labels. Our objective function aims to minimize the cross entropy between the target t and the prediction y. As for the optimization, we use the well-known backpropagation algorithm and Adam [4] to compute the parameters in SIL-CCNN.

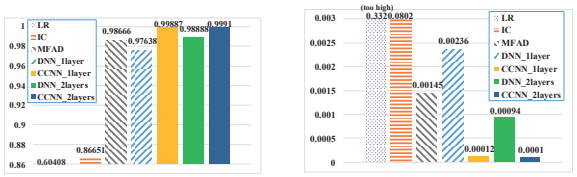
4. Experiments

In this section, we introduce experiments aiming at evaluating predicting performance in social networks using a synthetic dataset and a real-world dataset.



(a) Accuracy (b) MAE

Figure 3: Accuracy and MAE on the synthetic dataset of 5000 nodes



(a) Accuracy (b) MAE

Figure 4: Accuracy and MAE on twitter dataset

4.1 Dataset descriptions

Synthetic Data. Using the Lancichinetti-Fortunato-Radicchi (LFR) benchmark [5], we generate the synthetic dataset with 5000 nodes (2,666,674 traces). For the generation of diffusion data, we set the transmission probability uniformly between 0.1 and 1. We then choose a node at random to function as a source node and set it to be active.

Real Data. We crawl data from Twitter*1 for the period between September 2011 and May 2015. We use the experts in Healthcare Pundits*2 and Security*3 as the ini-

*1 https://twitter.com
*2 http://nursepractitionerdegree.org/top-50-health-care-pundits-worth-following-on-twitter.html
*3 http://www.marblesecurity.com/2013/11/20/100-security-

tial nodes, and crawl the followers of these experts to form a social network. The twitter data contains 20,453 users and 3,782,305 tweets. In this study, tweets were used as items, and the actions of sharing, replying, and liking are indications of the influence of a tweet.

4.2 Compared Methods

We include the following methods to predict the information diffusion results.

- **Logistic Regression** (LR). The in-degree and out-degree are included in a classifier for individual v.
- **Independent Cascade model** (IC). IC is a conventional diffusion model in which the probability of transmission from individual u to v is the ratio of items individual v adopted items from individual u.
- **Multiple Factor-Aware Diffusion model** (MFAD). MFAD is a diffusion model that can learn the social influence by multiple features [2].
- **SIL-DNN and SIL-CCNN**. We evaluated the performance of SIL-DNN and SIL-CCNN using one hidden layer (DNN_1layer, $CCNN\_1layer$) and two hidden layers ($DNN\_2layers$, $CCNN\_2layers$) in the following experiments.

4.3 Evaluated Metrics

We evaluate the performance of algorithms using the following two metrics.

- **Accuracy**. Accuracy is defined as the ratio of correct predicted answers over all answers in order to estimate the correctness of predictions.
- **Mean Absolute Error** (MAE). We computed the mean absolute error as follows: $MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$, where y_i is the truly adopted result of individual i, x_i is the estimated adoption probability of individual i, and n is the number of individuals in the network.

4.4 Results and Discussions

The comparison results on synthetic dataset is shown in Fig. 3. SIL-DNN and SIL-CCNN outperform the other three methods in the synthetic dataset. In addition, SIL-CCNN have higher accuracy and lower MAE than SIL-DNN using the same number of hidden layers. The results show that the information of the community structure actually helps the model to predict the influence results better. The proposed community-based CNN indeed extracts the local relations of individuals. Moreover, the performance of SIL-CCNN 2 layers is better than the SIL-CCNN 1 layer. Using more hidden layers also conducts a better result.

In the real twitter dataset, the results are shown in Fig. 4. SIL-DNN and SIL-CCNN still outperform the other three methods except for the DNN_1layer. We have examined the propagation results in these two datasets. We found that the influenced scale of the synthetic dataset is much larger than the Twitter dataset. This indicates that the diffusion results in synthetic dataset should be more difficult to predict. The superiority of SIL-CCNN over SIL-DNN shows that the performance can be improved by including the community information by the proposed community-based convolutional neural network.

experts-follow-twitter/

5. Conclusions and Future Works

In this work, we proposed two neural network architectures, SIL-DNN and SIL-CCNN, to identify social influences based on the propagation of information in a social network. The proposed framework makes it possible to obtain the diffusion results within the community. SIL-DNN and SIL-CCNN can predict the users who are actually influenced from the information without the Monte Carlo simulation. Experimental results demonstrate that SIL-DNN and SIL-CCNN both outperform existing methods.

For further improvement, we will design a strategy to extend the depth of SIL-CCNN in the future to overcome the problem with the insufficient number of traces. We also want to revise the structure of neural network to consider more features such as the content of items.

References

[1] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *In Proceedings of WWW*, pages 925–936, 2014.

[2] C.-K. Chou and M.-S. Chen. Multiple factors-aware diffusion in social networks. In *In Proceedings of PAKDD*, pages 70–81, 2015.

[3] D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *In Proceedings of h ACM SIGKDD*, pages 137–146, 2003.

[4] D. Kingma and J. B. Adam. A method for stochastic optimization. In *In Proceedings of ICLR*, 2015.

[5] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80, 2009.

[6] K. Saito, R. Nakano, and M. Kimura. Prediction of information diffusion probabilities for independent cascade model. In *In Proceedings of KES*, pages 67–75, 2008.

[7] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *In Proceedings of WSDM*, pages 643–652, 2012.

[8] L. Wang, S. Ermon, and J. E. Hopcroft. Feature-enhanced probabilistic models for diffusion network inference. In *In Proceedings of ECML-PKDD*, pages 499–514, 2012.