Accepted Manuscript

User behavior prediction via heterogeneous information preserving network embedding

Weiwei Yuan, Kangya He, Guangjie Han, Donghai Guan, Asad Masood Khattak

PII: S0167-739X(18)31567-X
DOI: https://doi.org/10.1016/j.future.2018.09.036
Reference: FUTURE 4471

To appear in: *Future Generation Computer Systems*

Received date : 1 July 2018
Revised date : 17 August 2018
Accepted date : 12 September 2018

Please cite this article as: W. Yuan, et al., User behavior prediction via heterogeneous information preserving network embedding, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.09.036

User Behavior Prediction via Heterogeneous Information Preserving Network Embedding

Weiwei Yuan[a,b], Kangya He[a], Guangjie Han[c,*], Donghai Guan[a,b], Asad Masood Khattak[d]

[a] *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*
[b] *Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210093, China*
[c] *Key Laboratory for Ubiquitous Network and Service Software of Liaoning province, School of Software, Dalian University of Technology, Dalian, 116024, China*
[d] *College of Technological Innovation, Zayed University, P.O. Box 144534. Abu Dhabi, United Arab Emirates*

Abstract

User behavior prediction with low-dimensional vectors generated by user network embedding models has been verified to be efficient and reliable in real applications. However, most user network embedding models utilize homogeneous properties to represent users, such as attributes or user network structure. Though some works try to combine the two kinds of properties, the existing works are still not enough to leverage the rich semantics of users. In this paper, we propose a novel heterogeneous information preserving user network embedding model, which is named HINE for user behavior classification in user network. HINE applies attributes, user network connection, user network structure, and user behavior label information for user representation in user network embedding. The embedded vectors considering these multi-type properties of users contribute to better user behavior classification performances. Experiments verified the superior performances of the proposed approach on real-world complex user network dataset.

Keywords: Heterogeneous information, network embedding, behavior

*Corresponding author
Email addresses: `yuanweiwei@nuaa.edu.cn` (Weiwei Yuan), `kangyahe@gmail.com` (Kangya He), `hanguangjie@gmail.com` (Guangjie Han), `dhguan@nuaa.edu.cn` (Donghai Guan), `Asad.Khattak@zu.ac.ae` (Asad Masood Khattak)

prediction, complex networks analysis

1. Introduction

User behavior prediction focuses on exploiting various information to infer the missing labels of unlabeled behavior [1, 2]. Predicting user behavior with the user network is a hot research topic nowadays. With the increasing of available information in user networks, the dimension of vectors describing users is getting larger. This makes user behavior prediction computational more expensive, especially for the large-scale networks [3]. Network embedding, which embeds users into a low-dimensional vector space by preserving user network structure or other user inherent properties, has therefore been proposed for use to facilitate more efficient user behavior prediction. With users represented with low-dimensional vectors, various vector-based machine learning algorithms can be more easily applied on user networks for further user behavior prediction analysis.

Existing works [4, 5, 6] have proposed various user network embedding models for user behavior prediction. The goal of user network embedding models is to minimize the reconstruction loss from the embedded vectors to the original vectors of users. This means the embedded matrix should maximize the contained information of the adjacency matrix representing the user network. Different kinds of information have been used for reconstruction loss evaluation: some models contain user network connection relationships between users [3, 7], some models contain user network structure information [5, 8], some models contain user attribute information [2, 9, 10, 11, 12], and some models contain two kinds of the information [6]. However, because of the diversity of user properties [13, 14], existing models, which consider only one or two user properties in reconstruction loss evaluation of user network embedding, are still not enough to leverage the rich semantics of users.

In this paper, we propose a novel heterogeneous information preserving user network embedding model, which is named HINE, for user behavior prediction with complex user network. Four kinds of heterogeneous information, including user attributes, user network connection, user network structure and user behavior label information, are contained for each user in the user network embedding. The reconstruction losses containing these four kinds of heterogeneous information are minimized simultaneously. User network connection and user network structure reflect the status of users in the user network, in which user network connection reflects the local user network relations and user network structure reflects global user network relations. User attributes reflect the inherent semantic properties of users, while user behavior label reflects the relationship between users according to these inherent semantic properties. Considering these heterogeneous information, the embedded vectors generated by HINE contribute to better user behavior prediction performances. Extensive experimental results validated the superiority of the proposed approach on user behavior prediction.

The rest of this paper is organized as follows: we first review some related work in Section 2, and then introduce the proposed approach in Section 3. Section 4 presents the experiments, followed by the conclusion in Section 5.

2. Related works

User behavior prediction is one of the typical applications of network embedding [1, 2]. By mapping vectors of users to a low-dimensional vector space via user network embedding, the efficiency of user behavior prediction can be greatly improved. In addition, since the main information of users is contained by user network embedding, user behavior prediction can also achieve reasonable prediction accuracy.

Based on the algorithms used for encoding the original vectors and decoding the embedded vectors, the existing user network embedding models can be categorized into matrix factorization based models and neural network based models [6]. Matrix factorization [4] based models use matrix factorization to map the original vectors into a low-dimensional vector space. In the matrix factorization based models, locally linear embedding model [15] uses the distance matrix of neighbors for the mapping of low-dimensional representation, spectral embedding model [16] uses the spectral radius decomposition of the Laplace matrix of a graph for the mapping of low-dimensional representation, sparse random projection model [17] uses the a random sparse matrix for the mapping of low-dimensional representation, and GraRep [18] carries out matrix factorization based on different sized random walk. Matrix factorization based models are time-consuming in large scale networks since matrix decomposition would be much more complex with the increasing scale of networks [6].

Neural network based models use neural networks to map the original vectors into a low-dimensional vector space [1]. In the neural network based models, Deepwalk [5] and node2vec [19] both use random walk to generate the sequence data like text, and then use the ID of users or text as the word of text to get the word vector by skip gram. The difference is that random walk held by Deepwalk is based on the weight of edges, while random walk held by node2vec is based on bias [19]; Line [3] uses KL divergence to measure the first-order similarity and second-order similarity between users, acting as the basis for the mapping of low-dimensional representation. Neural network based models have higher efficiency than matrix factorization based models in large scale networks [2, 1]. The proposed method belong to the category of neural network based models.

Existing models, regardless matrix factorization based models or neural network based models, only involve limited information describing users. The proposed method enriches the user representation by involving heterogeneous information.

3. The proposed method

Let $G = (V, E, \mathbf{X}, \mathbf{y})$ represent a user network, where V is a set of users,$n = |V|$, $E \subset V \times V$ is a set of links between users of V, $\mathbf{X} \subset \mathbb{R}^{n \times m}$ is the user attribute matrix of V , m is the dimension of user attributes, and $\mathbf{y} \subset \mathbb{R}^{n}$ is the user behavior label of ($\mathbf{y}_i$ is the label of v_i). Suppose the scale of the known user behavior labels is k, which are denoted as $L_i (1 \leq i \leq k)$. The unknown user behavior label is denoted as $L_{unlabeled}$ i.e., $\mathbf{y}_i \in L_1, L_2, ..., L_k, L_{unlabeled}$. To simplify the representation, the labeled user set and the unlabeled user set as denoted as $V_{labeled}$ and $V_{unlabeled}$ respectively, and $V = V_{labeled} \cup V_{unlabeled}$.

A user network embedding model can be denoted as:

$$\mathbf{h} = f_\theta(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is the original vector representing the user of the user network, $\mathbf{h}$ is the embedded vector of the user, $f_\theta(.)$ is the embedding algorithm, and denotes the parameters of the model.

HINE considers heterogeneous information for user representation. These information is used to describe the properties of users from different aspects. The reconstruction loss of embedded vectors should minimize the loss of these information simultaneously. The involved loss includes user attributes based loss, user network connection based loss, user network structure based loss, and user behavior label based loss, which are introduced in details as follows.

3.1. User attribute based loss

Since user attributes reflect the inner property of users in the user network, the attribute information of users is contained by HINE. User attribute based loss is the loss of user attribute information when reconstructing embedded vectors of users. This loss is measured by Deep Autoencoder in HINE. Deep Autoencoder can be divided into two parts: the encoder part and the decoder part [20, 21]. The intersection part of encoder and decoder, which is the central layer of Deep Autoencoder, is the compressed representation matrix learned by Deep Autoencoder.

Suppose the encoder part has N layers excluding the central representation layer, then the decoder part also has N layers excluding the central representation layer. Let $\mathbf{z}_i$ represent the i^{th} $(i = 1, 2, ..., N)$ layer of the encoder, let $\mathbf{W}_i$ represent the weight of $\mathbf{z}_i$, and let $\mathbf{b}_i$ represent the bias. Let $\mathbf{z}_i^{'}$ represent thei^{th} $(i = 1, 2, ..., N)$ layer of the decoder, let $\mathbf{W}_i^{'}$ represent the weight of $\mathbf{z}_i^{'}$, and let

$\mathbf{b}_i^{'}$ represent the bias. Let $\mathbf{x}$ represent the input matrix of Deep Autoencoder, then each layer of Deep Autoencoder is calculated as:

$$\mathbf{z}_1 = \tanh(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \tag{2}$$

$$\mathbf{z}_i = \tanh(\mathbf{W}_i\mathbf{z}_{i-1} + \mathbf{b}_i), i = 2, .., N \tag{3}$$

$$\mathbf{z}_1^{'} = \tanh(\mathbf{W}_1^{'}\mathbf{z}_N + \mathbf{b}_1^{'}) \tag{4}$$

$$\mathbf{z}_i^{'} = \tanh(\mathbf{W}_i^{'}\mathbf{z}_{i-1}^{'} + \mathbf{b}_i^{'}), i = 2, .., N \tag{5}$$

For the convenience of representation, the process shown in (2)-(5) is abbreviated as:

$$\mathbf{h} = f_{\mathbf{W},\mathbf{b}}(\mathbf{x}) \tag{6}$$

$$\hat{\mathbf{x}} = g_{\mathbf{W},\mathbf{b},\mathbf{W}',\mathbf{b}'}(\mathbf{x}) \tag{7}$$

where $\mathbf{h}$ is the compressed representation matrix learned by Deep Autoencoder, i.e., the matrix represented by the central layer; $\mathbf{W}$ and $\mathbf{b}$ are the weights and the bias of the encoder respectively; $\hat{\mathbf{x}}$ is the decoded matrix of $\mathbf{h}$ with the decoder, i.e., the matrix generated by Deep Autoencoder via the encoder and the decoder sequentially from $\mathbf{x}$; $\mathbf{W}^{'}$ and $\mathbf{b}^{'}$ are the weights and the bias of the decoder respectively. To reconstruct $\hat{\mathbf{x}}$ from $\mathbf{h}$, Deep Autoencoder should minimize the error between $\hat{\mathbf{x}}$ and $\mathbf{x}$.

The user attribute based loss is calculated as:

$$L_{att}(G) = \sum_{i=1}^{n} \|g_{\mathbf{W},\mathbf{b},\mathbf{W}',\mathbf{b}'}(\mathbf{X}_i) - \mathbf{X}_i\|_2 \tag{8}$$

where $g_{\mathbf{W},\mathbf{b},\mathbf{W}',\mathbf{b}'}(\mathbf{X}_i)$ is calculated by (7).

3.2. User network connection based loss

If two users are connected in the user network, it is more probably these two users have similar behavior [22], i.e., these two users are more probably to be classified in the same class by user behavior prediction. The user network connection information of users is therefore contained by HINE. User network connection based loss is the loss of user connection information when reconstructing embedded vectors of users.

To contain the user network connection in embedded vectors, if two users are connected in the user network, the distance between their embedded vectors should be minimized; otherwise, if two users are not connected in the user network, the distance between their embedded vectors should be maximized. The user network connection based loss is therefore calculated as:

$$L_{con}(G) = \sum_{i,j \in V} \mathbf{W}^{Con}_{i,j}(\|f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_i) - f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_j)\|_2 - \delta) \tag{9}$$

where δ is the threshold of user similarity, and $\mathbf{W}^{Con}_{i,j}$ is the weight of user network connection between v_i and v_j, which is calculated as:

$$\mathbf{W}^{Con}_{i,j} = \begin{cases} \alpha, if(v_i, v_j) \in E,\ or(v_i, v_j) \in E, \\ \beta, else \end{cases} \tag{10}$$

where α and β are hyper-parameters, and $\alpha,\ \beta \in (0, 1)$.

3.3. User network structure based loss

User network structure measures the global relationship of users in the user network, while user network connection measures the local relationship of users. Based on the user network structure, users can be categorized into different user network communities [23]. Users in the same user network community are more probably to be classified in the same class by user behavior prediction. The user network structure information of users is therefore contained by HINE. User network structure based loss is the loss of global relationship between users in the user network when reconstructing embedded vectors of users.

User network structure can be measured in different ways. HINE measures the multi-layer [24, 5, 25] structure of the user network. A multilayer user network [24] is a pair $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ of user networks, where $\mathcal{G} = \{G_\alpha | \alpha \in \{1, 2, ..., M\}\}$ is a family of graphs $G_\alpha = (V^\alpha, E^\alpha, \mathbf{X}^\alpha, \mathbf{y}^\alpha)$, G_α is the α^{th} layer of $\mathcal{M}$, M is the total number of layer, and $\mathcal{C} = V^k \times V^l$ ($k \neq l$ and $k, l \in \{1, 2, ..., M\}$) is the links connecting users of the k^{th} layer and users of the l^{th} layer. Suppose a user network $G = (V, E, \mathbf{X}, \mathbf{y})$ can be divided into M layers and $\mathcal{M}$ is its divided multilayer user network, then $V = V^1 \cup V^2 \cup ... \cup V^M$ and $E = C \cup E^1 \cup E^2 \cup ... \cup E^M$, and $\forall v_i = v_j^\alpha \; then \; \mathbf{y}_i = \mathbf{y}_j^\alpha$.

In a multilayer user network, users in the same layer of the user network are closely related, while users in different layers are loosely related or not related [24]. Since HINE is used for user behavior prediction task in this work, the relationships between users in a multilayer user network are measured by class similarities of users, i.e., users of the same layer should have similar class, while users of different layers should have different classes. Information entropy is used the measure each layer of the user network. Based on the Entropy principle of information theory, information entropy measures the purity of information in a class. The smaller entropy value a class has, the purer the users belongs to this class [26, 27]. Therefore, the smaller entropy value a layer has, it is more probably that users in this layer belong to the same class; the higher entropy value a layer has, it is less probably that users in this layer belong to the same class. An entropy based weight is defined for the similarity of user pair according to the user network structure:

$$\mathbf{W}_{i,j}^{Str} = Ent(V^k \cup V^l) \tag{11}$$

where $v_i \in V^k$, $v_j \in V^l$, V^k and V^l are the user set of the k^{th} layer and the l^{th} layer of $\mathcal{M}$, and $Ent(.)$ is the entropy of ., which is calculated as:

$$EntV^t = -\sum_{j=1}^{k} p_j \log p_j \tag{12}$$

where p_j is the probability of label $L_j$in the user set of the t^{th} layer V^t.

Considering the similarity of users structure, the user network structure based loss is defined as:

$$L_{str}(G) = \sum_{(i,j)\in E} \mathbf{W}_{i,j}^{Str}(\|f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_i) - f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_j)\|_2 \tag{13}$$

where $\mathbf{W}_{i,j}^{Str}$ is calculated by (11), and $f_{\mathbf{W},\mathbf{b}}(\cdot)$ is the embedded matrix of $(\cdot)$, which is calculated by (6).

3.4. User behavior label based loss

Existing user behavior labels in the training set help user behavior prediction to give better prediction on unknown user behavior labels. User behavior labels represent the relationship between users: if two users have the same user behavior label, these two users are closely related; while if two users have different user behavior labels, it is impossible for these users to have very close relationships [28]. The user relationships based on their user behavior labels are therefore contained by HINE. User behavior label based loss is the loss of use similarities according to their user behavior labels when reconstructing embedded vectors of users.

To contain user similarity according to user behavior labels, the distance between the embedded vectors of same-labeled users should be minimized, and the distance between the embedded vectors of different-labeled users should be maximized. For a user network $G = (V, E, \mathbf{X}, \mathbf{y})$, where $\mathbf{y}$ represents the user behavior labels of users, if $\mathbf{y}_i = L_{unlabeled}$, the user behavior label of v_i is unknown; otherwise, $\mathbf{y}_i \in \{L_1, L_2, ..., L_k\}$, the user behavior label of v_i is one of the k user behavior labels, and k is the total number of user behavior labels. The user behavior label based loss is calculated as:

$$\begin{aligned} L_{lab} =& \frac{1}{k}\sum_{i=1}^{k}(\frac{1}{|L_i|}\sum_{v_j\in Li}\|f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_j) - \frac{1}{|L_i|}\sum_{v_j\in L_i} f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_j)\|_2) \\ &- \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{l=i+1}^{k}\|\frac{1}{|L_i|}\sum_{v_j\in L_i} f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_j) - \frac{1}{|L_l|}\sum_{v_j\in L_l} f_{\mathbf{W},\mathbf{b}}(\mathbf{X}_j)\|_2 \end{aligned} \tag{14}$$

where $f_{\mathbf{W},\mathbf{b}}(\cdot)$ is the embedded matrix of $(\cdot)$, which is calculated by (6).

3.5. HINE model

HINE considers above losses simultaneously for user representation. For a user network $G = (V, E, \mathbf{X}, \mathbf{y})$, the optimal function of HINE is:

$$\min \alpha_1 L_{att}(G) + \alpha_2 L_{con}(G) + \alpha_3 L_{str}(G) + \alpha_4 L_{lab}(G) \tag{15}$$

where $\alpha_i (i = 1, 2, 3, 4)$ is the hyper parameter, which is used to adjust the influence of each loss; $L_{att}(G)$, $L_{con}(G)$, $L_{str}(G)$ and $L_{lab}(G)$ are calculated by (8), (9), (13) and (14) respectively.

Let denote θ any parameter of $\mathbf{W}$, $\mathbf{b}$,$\mathbf{W}^{'}$ and $\mathbf{b}^{'}$. Adam [29] is used to optimize these parameters in this work. To simplify the representation,$L(\theta) \triangleq \alpha_1 L_{att}(G) + \alpha_2 L_{con}(G) + \alpha_3 L_{str}(G) + \alpha_4 L_{lab}(G)$, the parameter optimization algorithm is represented as:

$$\hat{\theta} = Adam(\min_{\theta} L(\theta)) \tag{16}$$

where $\hat{\theta}$ is the optimized value of θ, and $Adam(\cdot)$ is the optimization function [29].

Algorithm 1 Parameter optimization of HINE

Input: $G = \{V, E, \mathbf{X}, \mathbf{y}\}$, the original user network.

Parameters: s, sampling ratio

Output: $\hat{\theta}$, optimized parameter

Initial

Sampling s edges from G, denote as $G^{'} = (V^{'}, E^{'}, \mathbf{X}^{'}, \mathbf{y}^{'})$

Sampling s edges from $G^{'}$, denote as $\overline{G^{'}} = (\overline{V^{'}}, \overline{E^{'}}, \overline{\mathbf{X}^{'}}, \overline{\mathbf{y}^{'}})$

Calculate $L_{att}(\overline{G^{'}})$, $L_{con}(\overline{G^{'}})$, $L_{str}(\overline{G^{'}})$ and $L_{lab}(\overline{G^{'}})$.

$\hat{\theta} = Adam(\min L(\theta))$

Return $\hat{\theta}$

The pseudo-code of parameter optimization is shown in Algorithm 1. The input of the algorithm is $G = (V, E, \mathbf{X}, \mathbf{y})$, and the output is the optimized parameters of HINE. Since the scale of the user network is usually very large in real applications, for the efficiency of calculation, a subgraph G'_s is extracted from G for the parameter optimization, in which s is the sampling ratio. Parameters are optimized until each loss is convergent by (16). With the optimized parameters, four kinds of loss, which are calculated by (8), (9), (13) and (14) respectively, are then merged to minimize the objective function of HINE.

4. Experimental results

Experiments are held on CiteSeer network dataset [6] to verify the performances of HINE on user behavior prediction. The dataset includes 4732 edges and 3312 users. Each user has a 3703-dimension user attribute vector. There are totally 6 user behavior labels in the dataset: Agents, AI, DB, IR, ML, and HCI. The distribution of these user behavior labels are given in Table 1. Line [3], Deepwalk [5], GraRep [18], node2vec [19] and Line [3] are used as the baseline method of user network embedding. Based on our experiences in the experiments, α_1 is set to be 0.01, α_2 and α_3 are set to be 2, and α_4 is set to be 4 in (15).

Table 1: The distribution of user behavior labels in experimental dataset.

User behavior label	Number	Ratio
Agents	596	18.00%
AI	249	7.52%
DB	701	21.17%
IR	668	20.07%
ML	590	17.81%
HCI	508	15.34%

Firstly, user behavior prediction performances with different embedding dimensions and different sampling ratios are measured, and the experimental results are given in Fig.1 and Fig.2 respectively. The embedding dimension is the dimension of each users embedded vector. The smaller the embedding dimension is, the higher efficiency the user network embedding model has. The sampling ratio is the ratio of data extracted from the original dataset to act as the training data. It relates to the density of the user network. The smaller the sample ratio is, the sparser the user network is. The classical classification model Logistic Regression [30] is used to classify each user with its embedded vector. When measuring the user behavior prediction performances with different embedding dimensions, the sampling ratio is set to be 30% in the experiments. When measuring the user behavior prediction performances with different sampling ratio, the embedding dimension is set to be 10 in the experiments. For each tested embedding dimension or sampling ratio, 10 times random experiments are held and the means of their user behavior prediction performances are used as the final user behavior prediction performances.

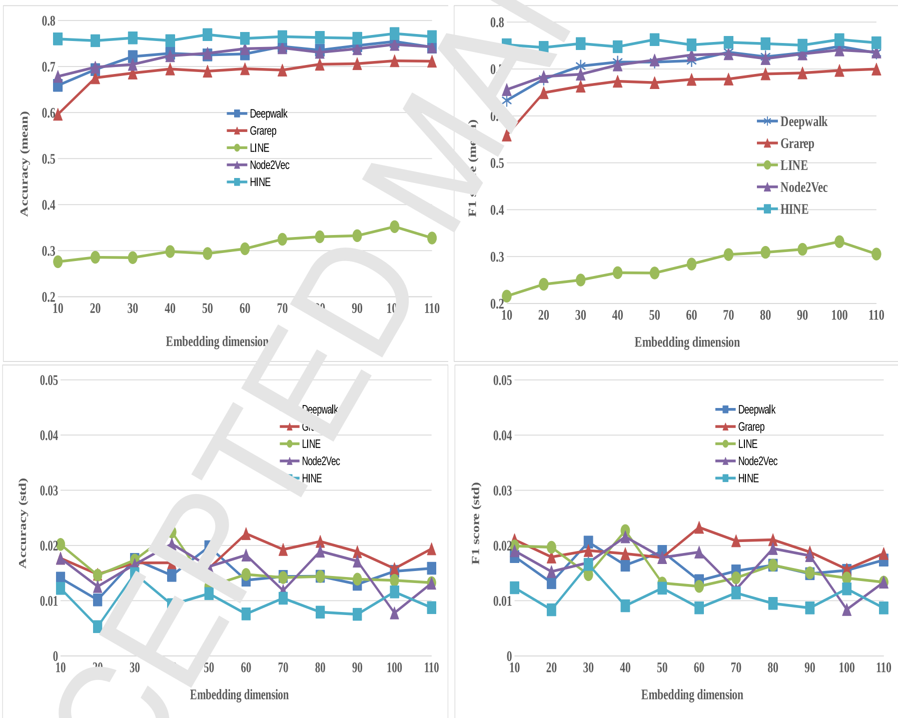


Figure 1: User behavior prediction performances with different embedding dimensions.

As shown in Fig.1 and Fig.2, HINE has superior user behavior prediction performances compared with baseline methods: classifying users with different

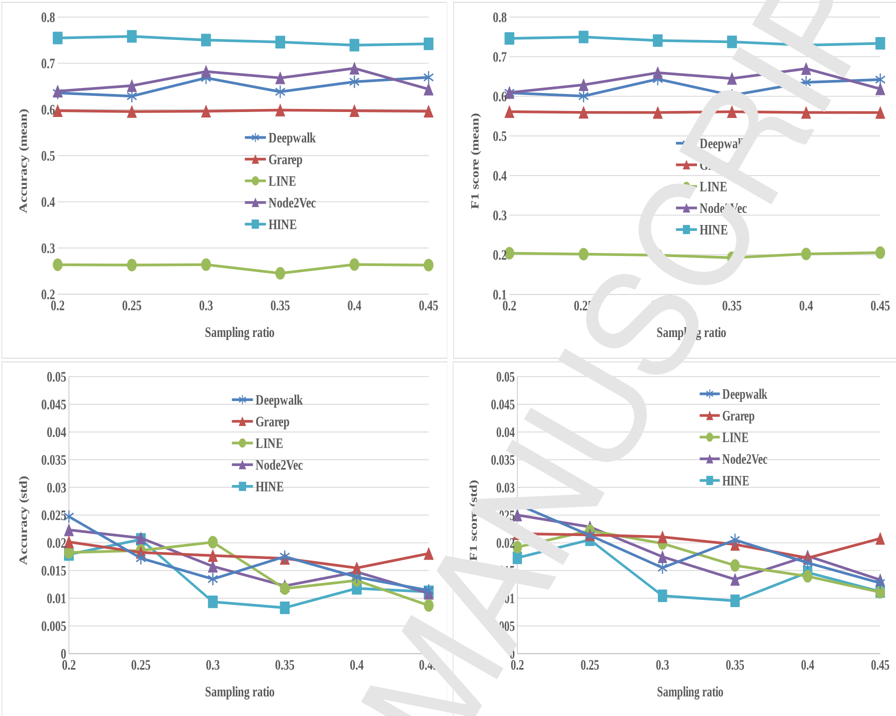


Figure 2: User behavior prediction performances with different sampling ratios.

embedding dimensions or sampling ratios, the mean of user behavior prediction accuracy and F1-score by using HINE are higher than using the baseline methods; the standard deviation of user behavior prediction accuracy and F1-score by using HINE are lower than using the baseline methods in majority cases. The user behavior prediction performances by using Line are the worst among all the tested user network embedding methods. The user behavior prediction performances by using Deepwalk and node2vec are similar, which are worse than using HINE. The user behavior prediction performances by using Grarep are better than Line, while significantly worse than HINE.

By using HINE, the user behavior prediction accuracy and F1-score only slightly change with different embedding dimensions or different sampling ratios. This means HINE is robust to the variety of embedding dimensions and sampling ratios. By using the baseline methods, the user behavior prediction

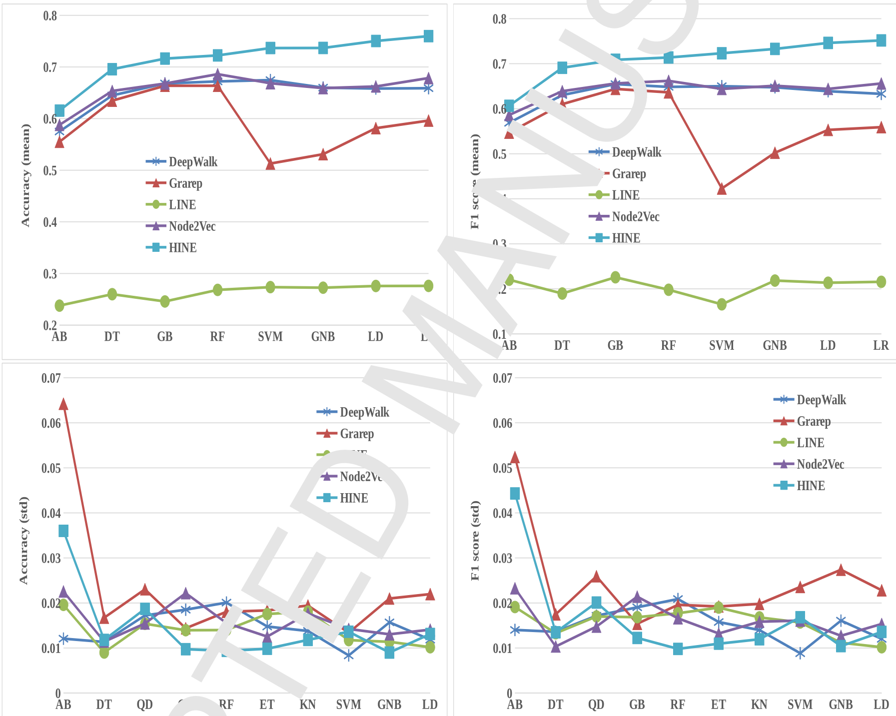


Figure 5: User behavior prediction performances with different classifiers.

accuracy and F1-score increase with the increase of the embedding dimension. The baseline methods do not have high user behavior prediction performances with low embedding dimensions, while even if embedding dimension is 10, HINE can still contribute to high user behavior prediction performances. This makes HINE to have much more advantages when embedding dimension tends to be low. Moreover, since HINE is not sensitive the sampling ratio, it is suitable to be used in the sparse user networks.

Secondly, user behavior prediction performances with different classifiers are measured, and the experimental results are given in Fig.3. To verify the performances of the proposed method with general classification algorithms, eight classical classifiers are used to classify users with the embedded vectors. These classifiers include AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), Linear Discriminant Analysis (LD) and Logistic Regression (LR). It is shown that the user behavior prediction performances by using HINE are significantly better than using the baseline methods with various classifiers. Classifying users with HINE, AB has the worst user behavior prediction performances, while LR has the best user behavior prediction performances. The user behavior prediction performances by using Deepwalk and node2vec are similar for different classifiers. And user behavior prediction performances by using these two methods are better than using other two baseline methods. The user behavior prediction performances by using Line are the worst among all tested user network embedding methods, regardless the classifiers. For AB, DT, GB and RF, using Grarep has similar user behavior prediction performances with using Deewalk or node2vec. While for SVM, GNB, LD and LR, user behavior prediction performances by using Grarep are much worse than using Deewalk or node2vec.

As mentioned above, the proposed method has superior performances compared with the existing methods. Generally speaking, the more information used in the network embedding, the better user behavior classification performances one can achieve. Existing works use limited information in their network embedding. For example, the structure information used in LINE only involves the first order neighbors and the second order neighbors, which is far less than the rich topology structural information. This leads to the worst performances of LINE among all methods. Deepwalk, Node2vec and Grarep contain the global structure information of the network via random walked node sequence. This leads to the similar user behavior classification performances by using these network embedding methods. Compared with the existing network embedding methods, HINE utilizes more information including the network connection information, the network community information, the node attribute information and the label information of nodes. These information reflects the relationships between users from different aspects, which contributes to the superior user behavior classification performances by using HINE.

5. Conclusion

This paper proposes a novel heterogeneous information preserving user network embedding model for user behavior prediction in user network. Heterogeneous information, including user attributes, user network connection, user network structure and user behavior label, is involved in reconstruction loss evaluation of embedded vectors. These information contains rich semantic properties of users from different aspects, reflecting the inherent properties of users and relationships between the target users and other users in user networks. Experimental results on open dataset demonstrate that the proposed approach is able to achieve higher user behavior prediction performances with various classifiers. Moreover, the proposed approach is suitable for user behavior prediction in sparse user network. It is also robust to the changing of embedding dimensions and can embed vectors with very small embedding dimensions accurately. In the future, we plan to involve more heterogeneous information in the proposed model for better user network embedding performances and user behavior prediction performances. Involving more information in HINE would increase the computational complexity. We will further design an optimized architecture for HINE which would preserve heterogeneous information more effectively. In addition, only very limited works focus on the network embedding of signed networks. This is because the signs of links make the network embedding much more complicate compared with the unsigned network[31]. In our future work, we will apply HINE into the signed networks and try to achieve better user behavior classification performances.

Acknowledgement

This research was supported by Nature Science Foundation of China (Grant No. 61672284), Natural Science Foundation of Jiangsu Province (Grant No. BK20171418), China Postdoctoral Science Foundation (Grant No. 2016M591841), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1601225C), the Fundamental Research Funds for the Central Universities (No.DUT17RC(3)094), Program for Liaoning Excellent Talents in University (No.LR2017009). This work was also supported by Zayed University Research Cluster Award # R18038.

References

[1] P. Goyal, E. Ferrara, Graph embedding techniques, applications, and performance: A survey, CoRR abs/1705.02801.

[2] W. L. Hamilton, R. Ying, J. Leskovec, Representation learning on graphs: Methods and applications, IEEE Data Eng. Bull. 40 (3) (2017) 52–74.

[3] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, LINE: large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015, 2015, pp. 1067–1077.

[4] D. Wang, P. Cui, W. Zhu, Structural deep network embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1225–1234.

[5] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: online learning of social representations, in: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, 2014, pp. 701–710.

[6] C. Li, S. Wang, D. Yang, Z. Li, Y. Yang, X. Zhang, J. Zhou, PPNE: property preserving network embedding, in: Database Systems for Advanced Applications - 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I, 2017, pp. 163–179.

[7] W. Yuan, K. He, D. Guan, L. Zhou, C. Li, Graph kernel based link prediction for signed social networks, Information Fusion 46 (2019) 1–10.

[8] G. Han, L. Liu, W. Zhang, S. Chan, A hierarchical jammed-area mapping service for ubiquitous communication in smart communities, IEEE Communications Magazine 56 (1) (2018) 92–98.

[9] X. Wang, T. Yang, M. Wortmann, P. Shi, F. Hattermann, A. Lobanova, V. Aich, Analysis of multi-dimensional hydrological alterations under climate change for four major river basins in different climate zones, Climatic Change 141 (2017) 1–16.

[10] T. Yang, T. Cui, C. Y. Xu, P. Ciais, P. Shi, Development of a new iha method for impact assessment of climate change on flow regime, Global & Planetary Change 156.

[11] G. Han, L. Zhou, H. Wang, W. Zhang, S. Chan, A source location protection protocol based on dynamic routing in wsns for the social internet of things, Future Generation Computer Systems.

[12] H. Wang, G. Han, C. Zhu, S. Chan, W. Zhang, Tcslp: A trace cost based source location privacy protection scheme in wsns for smart cities, Future Generation Computer Systems.

[13] G. Han, L. Liu, S. Chan, R. Yu, Y. Yang, Hysense: A hybrid mobile crowdsensing framework for sensing opportunities compensation under dynamic coverage constraint, IEEE Communications Magazine 55 (3) (2017) 93–99.

[14] G. Han, L. Zhou, H. Wang, W. Zhang, S. Chan, A source location protection protocol based on dynamic routing in wsns for the social internet of things, Future Generation Computer Systems 82 (2017) 689–697.

[15] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[16] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.

[17] P. Li, T. Hastie, K. W. Church, Very sparse random projections, in: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, 2006, pp. 287–296.

[18] S. Cao, W. Lu, Q. Xu, Grarep: Learning graph representations with global structural information, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015, 2015, pp. 891–900.

[19] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 855–864.

[20] G. E. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks., Science 313 (5786) (2006) 504–507.

[21] A. Jansen, G. Sell, V. Lyzinski, Scalable out-of-sample extension of graph embeddings using deep neural networks, Pattern Recognition Letters 94 (2017) 1–6.

[22] L. Liu, W. K. Cheung, X. Li, L. Liao, Aligning users across social networks using network embedding, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, 2016, pp. 1774–1780.

[23] G. Han, L. Liu, W. Zhang, S. Chan, A hierarchical jammed-area mapping service for ubiquitous communication in smart communities, IEEE Communications Magazine 56 (1) (2018) 92–98.

[24] M. Rosvall, C. T. Bergstrom, Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems, CoRR abs/1010.0431.

[25] M. Zitnik, J. Leskovec, Predicting multicellular function through multi-layer tissue networks, Bioinformatics 33 (14) (2017) i190–i198.

[26] I. V. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, Robust sound event classification using deep neural networks, IEEE/ACM Trans. Audio, Speech & Language Processing 23 (3) (2015) 540–552.

[27] M. Carpita, E. Ciavolino, A generalized maximum entropy estimator to simple linear measurement error model with a composite indicator, Adv. Data Analysis and Classification 11 (1) (2017) 139–158.

[28] X. Huang, J. Li, X. Hu, Label informed attributed network embedding, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, 2017, pp. 731–739.

[29] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.
URL http://arxiv.org/abs/1412.6980

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn:

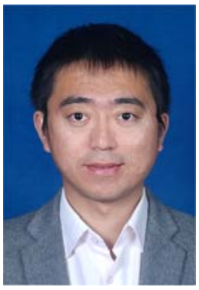
Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[31] W. Yuan, C. Li, G. Han, D. Guan, L. Zhou, K. He, Negative sign prediction for signed social networks, Future Generation Computer Systems.

Weiwei Yuan is an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. She received the B.S. and M.S. degrees from Harbin Engineering University, China, in 2002 and 2005, respectively, and the Ph.D. degree from the Department of Computer Engineering, Kyung Hee University, South Korea, in 2010. She has authored over 60 research papers in related international conferences and journals. Her current research interests include pattern recognition, social computing, and recommender systems.

Kangya He is studying master's degree in College of Computer Science and Technology from Nanjing University of Aeronautics and Astronautics. He received his B.S. in College of Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, China in 2016. His research interests are Machine Learning, Data Mining and Social Network Analysis.

Guangjie Han is currently a Professor with the School of Software, Dalian University of Technology, Dalian, China. He received the Ph.D. degree from Northeastern University, Shenyang, China, in 2004. From 2004 to 2006, he was a Product Manager for the ZTE Company. In February 2008, he finished his work as a Postdoctoral Researcher with the Department of Computer Science, Chonnam National University, Gwangju, Korea. From October 2010 to 2011, he was a Visiting Research Scholar with Osaka University, Suita, Japan. He is the author of over 220 papers published in related international conference proceedings and journals, and is the holder of 90 patents. His current research interests include sensor networks, computer communications, mobile cloud computing, and multimedia communication and security. Dr. Han has served as a Co-chair for more than 50 international conferences/workshops and as a Technical Program Committee member of more than 150 conferences. He has served on the Editorial Boards of up to 14 international journals, including the IEEE ACCESS, Telecommunication Systems, International Journal of Ad Hoc and Ubiquitous Computing, Journal of Internet Technology and KSII Transactions on Internet and Information Systems. He guest edited a number of special issues in IEEE Journals and Magazines. He has served as a Reviewer of more than 50 journals. He had been awarded the ComManTel 2014, ComComAP 2014, Chinacom 2014 and Qshine 2016 Best Paper Awards. He is a member of IEEE and ACM.



Donghai Guan is an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He received the Ph.D. degree from Kyung Hee University (KHU), Suwon, South Korea, in 2009. From 2009 to 2011, he was the Research Professor with the Department of Computer Engineering, KHU. From 2012 to 2014, he was an Assistant Professor with KHU. He has authored over 70 research papers in related international conferences and journals. His current research interests include machine learning, artificial intelligence, recommender system, social network analysis, and ubiquitous computing.

Asad Masood Khattak is an assistant professor at the College for Technological Innovation, Zayed University in Abu Dhabi. He received his M.S. in Information Technology from National University of Sciences and Technology, Islamabad, Pakistan in 2008. He got his Ph.D. degree in Computer Engineering from Kyung Hee University, South Korea in 2012. He worked as Post-Doctoral Fellow at Department of Computer Engineering, Kyung Hee University, South Korea for seven months and then he started working in the same college as Assistant Professor. He has authored/coauthored more than 50 journal and conference articles in highly reputed venues. His research interests include Knowledge Management, Context-aware Computing, Mobile Computing, Health Informatics and Forensics Analysis.

Highlights for this paper are listed as follows:

This paper proposes a novel heterogeneous information preserving user network embedding model, which is named HINE, for user behavior prediction. Four kinds of heterogeneous information, including user attributes, user network connection, user network structure, and user behavior label information, are used to represent each user in the user network embedding. Considering these heterogeneous information, the embedded vectors generated by HINE contribute to better user behavior prediction performances.