A link prediction approach based on deep learning for opportunistic sensor network

Jian Shu\textsuperscript{1,2}, Qifan Chen\textsuperscript{1,2}, Linlan Liu\textsuperscript{1,3} and Lei Xu\textsuperscript{1,2}

Abstract

Link prediction for opportunistic sensor network has been attracting more and more attention. However, the inherent dynamic nature of opportunistic sensor network makes it a challenging issue to ensure quality of service in opportunistic sensor network. In this article, a novel deep learning framework is proposed to predict links for opportunistic sensor network. The framework stacks the conditional restricted Boltzmann machine which models time series by appending connections from the past time steps. A similarity index based on time parameters is proposed to describe similarities between nodes. Through tuning learning rate layer-adaptively, reconstruction error of restricted Boltzmann machine goes stable rapidly so that the convergence time is shortened. The framework is verified by real data from INFOCOM set and MIT set. The results show that the framework can predict links of opportunistic sensor network effectively.

Keywords

Opportunistic sensor network, link prediction, similarity index, deep belief network, quality of service

Introduction

Opportunistic sensor network (OSN)\textsuperscript{1} is a new network evolved from the ad hoc network, which is a kind of self-organized network where the communication opportunities between nodes are generated by their movements. The data transmission of the OSN is realized by the store-carry-forward routing mechanism. This new networking mode has irreplaceable advantages in the non-fully connected networks which cannot execute the traditional multi-hop ad hoc networks protocol. Due to the easy development and low cost, OSN is widely applied in many fields such as vehicular ad hoc network, mobile data offloading,\textsuperscript{2} information sharing,\textsuperscript{3} and mobile computing.\textsuperscript{4}

With the appearance of critical, multimedia, and real-time applications, quality of service (QoS)\textsuperscript{5} in OSN is becoming increasingly significant. Due to link failures, node power failures, or different power management mechanisms, the dynamic topology of OSN makes it hard for underlying networks to ensure data transmission with QoS requirements. The key to solve this problem is capturing the change rule of the network topology and optimizing the routing algorithm through prediction methods. In this article, we propose a novel link prediction model of conditional deep belief network (CDBN) which models time series from the past time steps and a time-based similarity index which can describe the dynamic characteristics of OSN. The link prediction method based on deep learning is

Date received: 30 November 2016; accepted: 25 February 2017

Academic Editor: Fei Yu

Creative Commons CC-BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (http://www.creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (http://www.uk.sagepub.com/aboutus/openaccess.htm).
proposed to extract the features and study the change rule of the OSN topology, which provides significant support in designing routing algorithm and meets QoS requirements in OSN.

Related work

In recent years, as one of the research directions of data mining, link prediction has been widely studied by scholars in social networks and complex networks. Liu and Wu\(^9\) proposed an optimal probabilistic forwarding protocol to predict links. Zhou et al.\(^7\) studied a new similarity measure which is motivated by the resource allocation process on networks. Links can be simply predicted by this context based on prediction methods except some idealized assumptions which are not suitable for real networks. Liben-Nowell and Kleinberg\(^8\) utilized the similarity between nodes to predict links. Although the node attributes and network structure–based prediction method are efficient for static networks, it is not ideal for the dynamic OSN. By comparing several supervised machine learning algorithms, Al Hasan et al.\(^9\) found that the support vector machine (SVM) has a better performance in link prediction. Brouard et al.\(^10\) presented a new method for semi-supervised and transductive link prediction based on Output Kernel Regression, and the method shows that the semi-supervised learning algorithm is shown to be almost identical to the supervised learning algorithm in link prediction problems. The shallow learning algorithm is adopted in machine learning–based link prediction algorithms, which is insufficient in feature extraction for large-scale dynamic OSN.

With the concept of deep learning being put forward, neural network becomes a hot spot in the field of machine learning.\(^11\) Deep learning shows strong ability of learning essential feature of data sets from a few samples by learning a deep nonlinear neural network.\(^12\) In recent years, deep learning has made great achievements in image recognition,\(^13\) speech recognition,\(^14\) natural language processing,\(^15\) information retrieval, and other fields.

In this article, a deep learning based on link prediction model is established at first. Second, a novel similarity index is proposed according to the dynamic characteristics of OSN. Then, we optimize the learning rate of prediction model in order to make it more efficient. Finally, the experimental results show that the proposed link prediction model can predict links in OSN effectively.

Link prediction model of OSN

Restricted Boltzmann machine

Restricted Boltzmann machine (RBM)\(^16\) is a two-layer neural network model, which is composed of a visible layer and a hidden layer. There are several units among each layer, and the visible units represent the evaluation of objects, while the hidden units represent the state of objects. As it is shown in Figure 1, the hidden layer is h, the weight matrix is w, and the visible layer is v. The units are all connected between layers, while there are no units connected in each layer.

RBM defines \((v, h) \in \{0, 1\}^{N_v} \times \{0, 1\}^{N_h}\) as the distribution of the visible layer and the hidden layer. v and h mean that there are \(N_v\) visible units and \(N_h\) hidden units. The joint distribution function can be defined as follows

\[
P(v, h) = \frac{\exp(v^T W h + a^T v + b^T h)}{Z} \tag{1}
\]

\[
Z = \sum_{v, h} \exp(v^T W h + a^T v + b^T h) \tag{2}
\]

where \(W \in \mathbb{R}^{N_v \times N_h}\) is the weight matrix of layers and \(a\) and \(b\) are the biases of visible layer and hidden layer, respectively. Due to the non-neurons connected situation in each layer of RBM, the conditional distribution \(P(h|v)\) and \(P(v|h)\) can be defined as

\[
P(h_j = 1|v) = \sigma \left( b_j + \sum_i W_{ij} v_i \right) \tag{3}
\]

\[
P(v_i = 1|h) = \sigma \left( a_i + \sum_j W_{ji} h_j \right) \tag{4}
\]

where \(\sigma\) is the logistic function, which is defined as \(\sigma(z) = 1/(1 + \exp(-z))\).

In all models based on RBM, the normalized factor \(Z\) makes it difficult to solve maximum likelihood value. However, gradient can be evaluated by contrastive divergence algorithm which is proposed by Hinton.\(^17\) This approximate method makes the updating rules of gradient more simple, the parameters of the model are defined as follows

\[
\Delta W_{ij} \propto <v_i h_j>_{data} - <v_i h_j>_{recon} \tag{5}
\]

\[
\Delta a_i \propto <v_i>_{data} - <v_i>_{recon} \tag{6}
\]
The contrastive divergence algorithm can also be used in CRBM to update the weight and the static biases where they have the same form as equations (5)–(7) but have a different effect because the states of the hidden units are now influenced by the previous visible units. The gradients are now summed over all time steps

\[
\Delta W_{ij} \approx \sum_t \langle v_i h_j, t \rangle_{data} - \langle v_i h_j, t \rangle_{recon} \tag{12}
\]

\[
\Delta A_{ki} \approx \sum_t \langle v_i v_k, t \rangle_{data} - \langle v_i v_k, t \rangle_{recon} \tag{13}
\]

\[
\Delta B_{kj} \approx \sum_t \langle h_j v_k, t \rangle_{data} - \langle h_j v_k, t \rangle_{recon} \tag{14}
\]

\[
\Delta a_i \approx \sum_t \langle v_i, t \rangle_{data} - \langle v_i, t \rangle_{recon} \tag{15}
\]

\[
\Delta h_j \approx \sum_t \langle h_j, t \rangle_{data} - \langle h_j, t \rangle_{recon} \tag{16}
\]

where \( \langle \cdot \rangle_{data} \) is an expectation of data distribution and \( \langle \cdot \rangle_{recon} \) is the K-step reconstruction distribution as obtained by alternating Gibbs sampling, starting with the visible units clamped to the training data.

While learning a CRBM, there is no need to proceed sequentially through the training data sequences. The updates are only conditional on the past \( N \) time steps, not the entire sequence. As long as we isolate “chunks” of \( N + 1 \) frames (the size depending on the order of the directed connections), these small windows can be mixed and formed into mini-batches. To speed up the learning, we assemble these chunks of frames into “balanced” mini-batches of size 50.

**CDBN**

Deep belief network (DBN) is a deep learning neural network composed by multi-layer RBM which is proposed by Hinton.\(^{19}\) In this article, the link prediction adopted in the OSN uses the same method as mentioned above. Figure 3 shows a CDBN which is stacked by several CRBMs. The steps to predict links in OSN with CDBN can be summarized as follows: first, multi-layer CRBM is used to extract intrinsic characteristics from the dynamic network according to the meeting records among the snapshots. After training the link prediction model, we connect a softmax classifier to the topmost hidden layer of the CDBN in order to verify the
accuracy of the prediction model by test data set. However, there are two problems need to be taken into account:

(1) OSN is a dynamic network with frequent changes over time, so the link prediction indexes which applied to social network are not suitable for OSN. So, we need to establish the indexes to describe the dynamic characteristics of OSN.
(2) The learning rate is a hot spot in the research of CDBN. If the learning rate is too small, the training time will be too long. On the contrary, it would lead to the instability of the prediction model.

**Link prediction index based on time parameter**

In social network, Zhou et al.\(^7\) proposed the Local Path (LP) similarity index to predict links. The index takes third-order path into account which is defined as

\[
S^{LP} = A^2 + \alpha \cdot A^3
\]  

(17)

where \(\alpha\) is an adjustable parameter, \(A\) is the adjacency matrix of the network. \((A^3)_{xy}\) is the number of paths between \(v_x\) and \(v_y\), and the path length is 3.

All the communication nodes must be considered in OSN but the LP index ignores the one-order path parameter, so the one-order path is added into LP index and it can be defined as

\[
S^{LP'} = A + \alpha \cdot A^2 + \alpha^2 \cdot A^3
\]  

(18)

In OSN, we assume that there is a node pair \((x, y)\) to be predicted. We can find that link prediction is influenced by the time-related factors due to the dynamics of OSN, which is defined as follows:

\(t_1\): recent contact time with the target node.
\(t_2\): total length of time connected to the target node.
\(f\): frequency of contact with the target node.

With the analysis above, the time parameter \(TP\) can be defined as

\[
TP = t_1 + t_2 + t_3
\]  

(19)

It is assumed that the start time of time period \(T\) is \(T_1\) and the end time is \(T_2\). The index of recent contact time is defined as \(t_1 = (t_1 - T_1)/(T_2 - T_1)\). The index of relative contact time is defined as \(t_2 = t_2/(T_2 - T_1)\) and the contact frequency of node pair \((x, y)\) is defined as \(t_3 = fD(x)\) where \(D(x)\) is the total degree of nodes in time \(T\), which is defined as \(D(x) = D_{T_1}(x) + D_{T_2+1}(x) + \cdots + D_{T_2}(x)\).

The weight matrix of time parameter \(W_{tp}\) can be calculated by \(TP\). LP similarity index based on time parameter can be obtained by combining equation (18) with \(W_{tp}\), which is defined as

\[
S^{TPLP} = W_{tp} + \alpha \cdot W_{tp}^2 + \alpha^2 \cdot W_{tp}^3
\]  

(20)

By substituting the adjacency matrix of the network with the time parameter matrix, the similarity index of link prediction can be described from structure and time-varying property.

**Layer-adaptive learning rate**

The selection of learning rate is very important in the process of CRBM training. It tends to empirically choose a smaller learning rate to ensure the stability of the system. In this article, the relation between the gradient of the single layer and the learning rate is taken into account. The gradient of the current layer and the value of the global learning rate determine the value of the current layer learning rate in the same layer. The learning rate \(t_l^{(k)}\) is defined as follows

\[
t_l^{(k)} = \hat{t}^{(k)} \cdot \left(1 + \log \left(1 + \frac{1}{\left(\|g_l^{(k)}\|_2\right)}\right)\right)
\]  

(21)

where \(\hat{t}^{(k)}\) is the global learning rate at the Kth iteration and \(t_l^{(k)}\) is the adaptive learning rate at the Kth iteration in layer \(l\) and it is related to the gradient \(g_l^{(k)}\) of the current iteration of the current layer.

**Experiments and analysis**

**Experimental data**

This section will adopt INFOCOM05 and MIT to carry on the simulation experiment and verify the effectiveness of the CDBN model.
The reference information of the two experimental data sets is shown in Table 1 (in CRAWDAD, there are more details about INFOCOM05 and MIT data sets).

| Data set     | INFOCOM05 | MIT |
|--------------|-----------|-----|
| Device       | iMote     | Phone |
| Mobile nodes | 41        | 97   |
| Duration(days)| 3 days    | 246 days |
| Network type | 120       | 300  |
| Sampling     | 120       | 300  |
| Intervals    |           |      |
| Total contacts | 227,657  | 285,512 |

Table 2. AUC of similarity indexes.

| Index | MIT       | INFOCOM05 |
|-------|-----------|-----------|
| CN    | 0.7045    | 0.7318    |
| AA    | 0.7136    | 0.7415    |
| LP    | 0.7818    | 0.8291    |
| Katz  | 0.7727    | 0.8433    |
| TLP   | 0.8011    | 0.8543    |

AUC: area under the curve; CN: Common Neighbors; AA: Adamic-Adar; LP: Local Path.

Experimental results and analysis

First, in this section, the tradition link prediction index and the LP similarity index based on the time are compared to validate the effectiveness of the similarity index. Then, the CDBN model is constructed to compare with the results of the link prediction with two data sets. The influence of the single-layer adaptive learning rate and the sample dimension on the prediction with the proposed CDBN is analyzed in the final.

In the experiment, the similarity index which proposed by equation (20) is compared with Common Neighbors (CN) index, Adamic-Adar (AA) index, LP index, and the Katz index, the area under the ROC curve (AUC) values of each index are shown in Table 2.

From Table 2, it can be obtained from the results that the CDBN can improve the prediction accuracy further on the basis of TLP index.

From Figure 5, when the initial learning rate is 0.01, after 20 iterations of the CRBM, the error of RBM reconstruction under the adaptive learning rate of single layer reaches a relatively stable state. But under the fixed learning rate, the number of iterations of CRBM is nearly 40. Therefore, the adaptive learning rate of the single layer can improve the rate of convergence of CRBM networks when it compares with fixed learning rate. It can also improve the computational efficiency of CDBN.

In the experiment, in order to find the most suitable sample dimension of CDBN, the average accuracy of CDBN in different dimensions of the sample is obtained through calculating the 20 pairs of nodes under two data sets.

From Figure 6, when the sample dimensions are 75 or 100, CDBN will own a better average accuracy under the MIT data set and the INFOCOM data set.

In this experiment, we set the sample dimension to 75 for INFOCOM data set and 100 for MIT data set. The parameters of CDBN are the same as the first experiment. With the optimal parameters, we calculate with Table 2, it can be obtained from the results that the CDBN can improve the prediction accuracy further on the basis of TLP index.
the average AUC of 20 pairs of nodes in INFOCOM data set and MIT data set, respectively.

From Table 3, it can be seen that the average AUCs of CDBN with different similarity indexes are significantly improved when compared with the static similarity indexes in Table 2.

### Conclusion

In this article, a CDBN model based on the deep learning is proposed to achieve the link prediction of OSNs. By combining the CDBN with proposed similarity index, the change law of OSN can be captured accurately while improving the QoS in OSN.

The experimental results show that the links in OSN can be predicted better with the help of our time-based similarity index and the prediction model of CDBN.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by grants from the National Natural Science Foundation of China (nos 61262020, 61363015, 61501218, and 61501217).

### References

1. Spyropoulos T, Psounis CS and Raghavendra CS. Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In: Proceedings of the 2005 ACM SIGCOMM workshop on delay-tolerant networking, Philadelphia, PA, 26 August 2015, pp.252–259. New York: ACM.
2. Han B, Hui P, Kumar VSA, et al. Mobile data offloading through opportunistic communications and social participation. IEEE T Mobile Comput 2011; 11(5): 821–834.
3. Jung S, Lee U, Chang A, et al. Bluetorrent: cooperative content sharing for bluetooth users. Pervasive Mob Comput 2007; 3(6): 609–634.
4. Li Y and Wang W. Can mobile cloudlets support mobile applications? In: Proceedings of the IEEE conference on computer communications (INFOCOM 2014), Toronto, ON, Canada, 27 April–2 May 2014, pp.1060–1068. New York: IEEE.
5. Tanase M and Cristea V. Quality of service in large scale mobile distributed systems based on opportunistic networks. In: Proceedings of the IEEE workshops of international conference on advanced information networking and applications, Singapore, 22–25 March, pp.849–854. New York: IEEE.
6. Liu C and Wu J. An optimal probabilistic forwarding protocol in delay tolerant networks. In: Proceedings of the ACM international symposium on mobile ad hoc networking and computing (MOBIHOC 2009), New Orleans, LA, 18–21 May 2009, pp.105–114. New York: ACM.
7. Zhou T, Lü L and Zhang YC. Predicting missing links via local information. Eur Phys J B 2009; 71(4): 623–630.
8. Liben-Nowell D and Kleinberg J. The link prediction problem for social networks. J Assoc Inf Sci Technol 2007; 54(7): 1345–1347.
9. Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning. Proc SDM Workshop Link Anal Counterterror Secur 2005; 30(9): 798–805.
10. Brouard C, D’Alch-Buc F and Szafranski M. Semi-supervised penalized output kernel regression for link prediction. In: Proceedings of the international conference on machine learning (ICML 2011), Bellevue, WA, 28 June–2 July 2011, pp.593–600.
11. Hinton BGE, Osindero S and Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput 2006: 18: 1527–1554.
12. Bengio Y and Delalleau O. On the expressive power of deep architectures. In: Proceedings of the international conference on algorithmic learning theory, Espoo, 5–7 October 2011, pp.18–36. Berlin: Springer.
13. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vision 2015; 115(3): 211–252.
14. Dahl GE, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE T Audio Speech 2012; 20(1): 30–42.
15. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. *J Mach Learn Res* 2011; 12(1): 2493–2537.
16. Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006; 313(5786): 504–507.
17. Hinton GE. Training products of experts by minimizing contrastive divergence. *Neural Comput* 2002; 14(8): 1771–1800.
18. Palmqvist S, Sderfeldt B and Arnbjerg D. Modeling human motion using binary latent variables. In: *Proceedings of the conference on advances in neural information processing systems*, Vancouver, B.C., Canada, 4–7 December 2006, pp.1345–1352. Cambridge, MA: MIT Press.
19. Hinton G. *Deep belief nets*. New York: Springer, 2011.
20. Kotz D and Henderson T. Crawdad: a community resource for archiving wireless data at Dartmouth. *IEEE Pervas Comput* 2005; 4(4): 12–14.
21. Chaturvedi A. *Link prediction analysis in social networks*. Saarbrücken, Germany: LAMBERT Academic Publishing (LAP), 2013.
22. Adamic LA and Adar E. Friends and neighbors on the web. *Soc Networks* 2003; 25(3): 211–230.
23. Katz L. A new status index derived from sociometric analysis. *Psychometrika* 1953; 18: 39–43.