A method based on compressive sensing to detect community structure using deep belief network

Liangliang Zhang, Haijia Wu, Jing Feng, Xiongwei Zhang
PLA University of Science and Technology
Nanjing, China
vermouthlove@hotmail.com, wu_haijia@163.com

***Abstract*—A deep learning scheme based on compressive sensing to detect community structure of large-scale social network is presented. Our contributions in this work are as follows: First, we reduced the high-dimensional feature of social media data via compressive sensing by using random measurement matrix; Second, deep belief network is employed to learn unsupervised from the low-dimensional samples; Finally the model is fine-tuned by supervised learning from a small scale sample sets with class labels. The effectiveness of the proposed scheme is confirmed by the experiment results.**

Keywords—compressive sensing; community structure; social network; deep belief network(DBN)

I. Introduction

Over the past decade, research on community detection has mostly focused on small-scale social network. These algorithms can approximately divide into three categories: The first one is based on graph theory including GN [1] and FastGN[2]. Some other methods have been developed based on matrix factorization. For example, symmetric nonnegative matrix (SymNMF) has promising performance [3]. The methods based on optimization called N-Cut and A-Cut were proposed in [4].These methods can help us understand the network structure and reveal the network functions and have high efficiency when the size of network is small. However, large-scale and high-dimensional network brings low efficiency and the complexity of these algorithms grows exponentially.

Compressive sensing (CS) [5][6] has outstanding performance in processing sparse data. This theory has proven to break the Nyquist sampling theorem in data sampling process and is able to reconstruct the initial signal accurately from the fewer projections by using the sparse priors of it. Since the characteristic of social network is high-dimension and large scale, the connections between nodes is relatively less. For example, the sparsity of MovieLens dataset is 4.5%, the sparsity of Netflix is 1.2%, and the sparsity of Delicious is 0.046%. Taobao has eight hundred millions commodities at present, however, the number of products that an average Taobao user browsed is less than 1000, so its sparsity is below one millionth. To the best of our knowledge, the scale of social network is larger, the dataset is sparser. CS is capable of reducing dimension of dataset to extract the essential information by utilizing random measurement matrix, so it is suitable to handle high-dimensional network.

Deep learning (DL) [7][8] is a rising algorithm of multilayer neural network. This method solves the problem of local minimum and unsupervised training. DL adopts unsupervised learning algorithm in training phase**,** and utilize small labeled samples to supervised fine-tune the model in the final phase. The representative set is selected from large-scale social network, and its community labels can be obtained through traditional algorithms mentioned above.

In order to learn community structure of large-scale and high-dimensional network, we present a deep learning scheme to detect community structure based on compressive sensing in this paper. The proposed approach can improve the efficiency of extracting community structure process from two aspects. On the one hand, by using random measurement of compressive sensing theory, we can reduce the dimension of social network feature. On the other hand, by using deep belief network (DBN)[9] model, the learning and prediction problem of large-scale community structure become feasible.

II. Four Steps of Proposed Scheme

The overall procedure of the proposed community detection scheme is shown in Fig. 1. The scheme has four steps: Firstly, community structure of the representative set which extracted from original adjacency matrix is detected via GN algorithm. Then, the dimension of original adjacency matrix is reduced via random measurement matrix. Unsupervised training of DBN is implemented by utilizing low-dimensional feature samples. Finally, the small set of labeled samples that obtained in the first step is used to fine-tune the DBN with supervision.

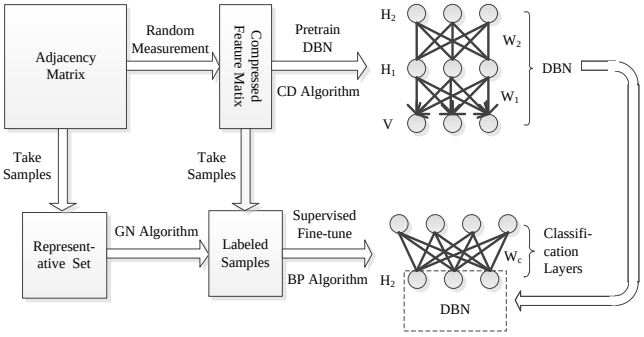

Fig.1. Overall procedure of the algorithm

A. Detect Community Structure of Representative Set

The representative set is a number of nodes $\tilde{V}$ that randomly selected from the large-scale complex networks V by keeping the connections between nodes. The connections are transformed into adjacency matrix A and adjacency matrix of representative set is $\tilde{A}$.

GN is a traditional community detection algorithm. The fundament thought of GN is that connection strength in one community is high and is usually low between communities. The algorithm employs edge betweenness to determine the attribute of the edges. Edge betweenness of an edge is defined as the number of shortest paths between pairs of vertices that run along it. Thus, the edge betweenness is high if the edge connecting communities. By considering this, the communities can be divided though deleting these edges. In basis of calculating edge betweenness, GN repeated iteration to reveal community structure.

The representative set is denote by the triple $S = \{< \tilde{v}_i, c_j, s_{ij} > | \tilde{v}_i \in \tilde{V}, c_j \in \mathrm{C}, s_{ij} = 0 \, or \, 1\}$, where C is community labels set, and the number of elements in C is N, s_{ij} is the tag that if $\tilde{v}_i$ belongs to c_j, s_{ij}=1 else s_{ij}=0.

B. Reduce Dimensions of Adjacency Matrix

Candes and Donoho proposed CS theory in [5][6]. The theory is to reconstruct the original signal X when the measurement matrix $\Phi \in R^{M \times N} (M \square N)$ and linear measurement value $Y \in R^M$ are known, and that can be written as

$$Y = \Phi X \tag{1}$$

Obviously, (1) has infinite solutions. CS theory indicates that X can be reconstructed accurately by solving the minimum norm of l_0. There are two requirements to confirm the reconstruction: a) X must be sufficient sparse. b) Measurement matrix Φ satisfies restricted isometry property (RIP) [10]. The expression is formulated as

$$\hat{X} = \arg\min \|X\|_0 \quad s.t. \ \Phi X = Y \tag{2}$$

By designing appropriate measurement matrix, high-dimensional feature is reduced without losing information. Bernoulli random measurement matrix is adopted in this paper, which is given as

$$\Phi \in R^{M \times N} : \Phi(i, j) = \frac{1}{\sqrt{M}} g_{i,j}, \quad g_{i,j} \sim \begin{pmatrix} 1 & -1 \\ 0.5 & 0.5 \end{pmatrix} \tag{3}$$

The RIP of Bernoulli random measurement matrix is demonstrated in [11].

CS is capable of providing dimensional-controlled training samples for DBN though reducing dimension of dataset without losing feature information. By adopting Bernoulli random measurement matrix M, dimension of adjacency matrix can be reduced, which is given as $A_c = A \times M$. Here, A_c means the feature of the compressive samples. The state space of samples inputting into DBN is {0, 1} or real number between zero and one. Thus, the elements in A_c must be normalized according to dimension. We focused on effect of two kinds of normalization method on deep learning community structure. Remapping is a way to distinguish the various types of samples at maximum and map the feature of all samples between zero and one according to dimension. The features satisfied even distribution. Another way is identically distributed normalization. This method compresses the dimension between zero and one directly and keeps the original distribution. By applying two normalization methods, the modularity of community structure is improved.

C. Train Deep Belief Network without Supervision

Hilton proposed DBN in [7]. Compared with traditional neural networks, DBN has a lot of advantages. Firstly, it's a kind of unsupervised learning algorithm which is suitable for processing big data without labels; secondly, the training algorithm of DBN, which is called Contrastive Divergence (CD), is more efficient than Back Propagation (BP) algorithm.

As shown in Fig. 2, DBN is composed of stacked Restricted Boltzmann Machines (RBM) which is a generative stochastic neural network that can learn a probability distribution over its set of inputs. As its name implies, RBM is a variant of Boltzmann machine, with the restriction that the neurons must form a bipartite graph: it has a visible layer V, corresponding to features of the inputs, and a hidden layer H that are trained, and each connection in an RBM must connect a visible unit to a hidden unit. We define the structure of DBN as $[\mathrm{n}_1, \mathrm{n}_2, \cdots, \mathrm{n}_k]$, where k is the number of RBM layers, n_1 is the number of input layer nodes, $\mathrm{n}_i (2 \le i \le k)$ is the number of hidden layers nodes, $[W_1, W_2, \cdots, W_{k-1}]$ means the transfer matrix between each layer.

Each layer of DBN is unsupervised trained by CD algorithm. Training set is A_c, n_1 is equal with the dimension of A_c. Input samples are mapped into different feature spaces by applying CD algorithm to keep feature information as much as possible. This training process is regarded as initializing weight parameters of multilayer BP neural network. Moreover, DBN overcomes local optimum and slow convergence speed via initializing.

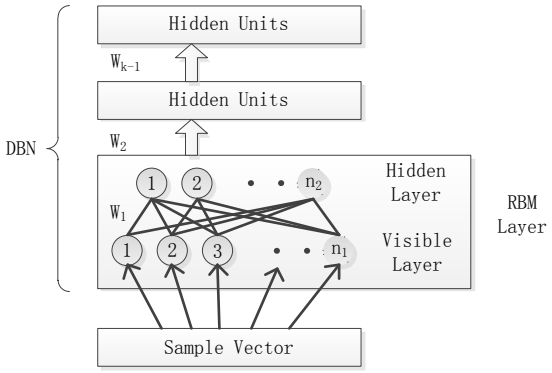

Fig.2. Training procedure of DBN

D. Fine-tune the Model with Supervision

After training DBN model, DBN build BP neural network for every class of community. BP neural network estimate samples whether belongs to the community via outputting zero or one. The procedure of the proposed fine-tune scheme is shown in Fig. 3. RBM only guarantee weight of own layer optimal mapping, however, it cannot achieve optimal for whole model. Because of this, model is fine-tuned though error back propagation. Though the training process, models N is generated.

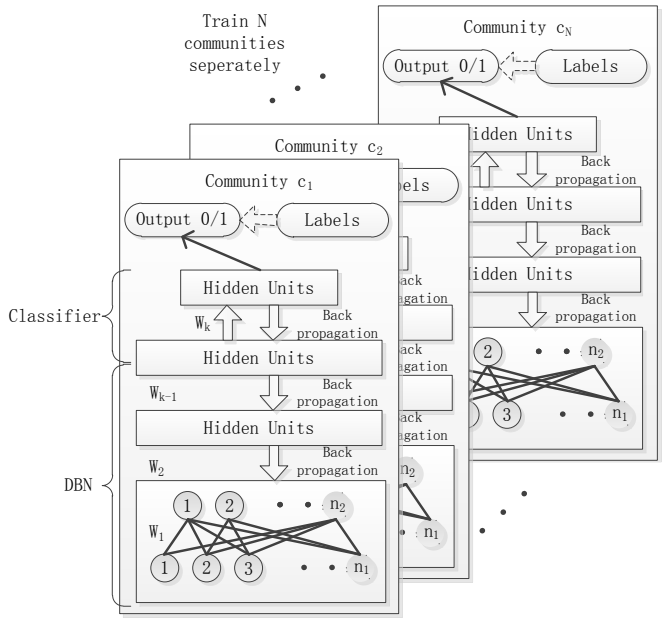

Fig.3. Fine-tuning procedure of the model

Recognition model of community structure is formed which can detect community via adjacency relationship. Recognition steps are as follows: First, vectors of adjacency matrix are compressed by utilizing random measurement matrix. Second, we input compressive feature vectors to the model. Finally, model recognizes which communities they are belong to.

III. Experiments and Results

The modularity [12] is used to judge the performance of the detected community structure. The modularity Q represents the strength of communities, which is given as

$$Q = \sum_{i=1}^{N} \left[\frac{l_{c_i}}{L} - \left(\frac{d_{c_i}}{2L} \right)^2 \right] \tag{4}$$

The performance of the two kinds of normalization methods is compared in this paper. Test data set is power[1] with 4914 nodes, and the sparsity of the data set is 0.054%. We extract first 1000 nodes to constitute the representative set. Community structure of representative set is detected via GN algorithm and we divide 33 communities, value of Q is 0.7232. Then, we set 3 layers of DBN, number of input layer is n_1, 2 hidden layers are respectively 400 and 200. Moreover, number of classification layer nodes is 100. By compressing power dataset with Bernoulli random measurement matrix, we can obtain the feature dimension n_1.

Table 1 shows the result of two kinds of normalization methods under different samples feature dimension. As depicted in Table 1, remapping normalization has better performance on community detection compared with identically distributed normalization. $n_1 = 1000$, then Q=0.2914, when using remapping normalization. On the contrary, if we increase n_1, the performance is poor and the best value is $n_1 = 800$. We can infer that higher feature dimension can only maintain original feature information with leading to more parameters. In addition, overfitting is generated because of the increasing parameters and fixed samples. Remapping adjusts feature distribution to improve discrimination of samples. It is proved that remapping is more suitable to detect community structure.

TABLE I. PERFORMANCES OF COMMUNITY STRUCTURE DETECTION UNDER TWO NORMALIZATION METHODS

Normalization Methods	Modularity Q of detected community structure					
	n_1 =200	400	600	800	1000	1200
Remapping	0.0873	0.1711	0.2269	0.2821	0.2914	0.2227
Identically distributed	0.0045	0.0419	0.0477	0.0480	0.0240	0.0164

The structure of DBN has significant influence on community detection. As influence measures, we adopt incremental hidden layers such as setting 1000 input nodes, 100 classification layer nodes. The representative set power is compressed to 1000.

First we test the performance of single hidden layer DBN. As showed in Table 2, the modularity of community structure is highest (Q=0.2761) since $h = 1$ and $n_2 = 400$. Then, we set first hidden layer that of $n_2 = 400$ and increase the number of second hidden layer. The modularity of community structure is highest (Q=0.2914) since $h = 2$ and $n_3 = 200$. However, we set second hidden layer that of $n_3 = 200$ and increase the number of third hidden layer. The modularity of community structure is highest (Q=0.2010) since $h = 3$ and $n_4 = 200$. The performance degrade due to overfitting that generating more model parameters and requiring more samples for training.

TABLE II. EFFECT OF DIFFERENT DBN STRUCTURES TO THE PERFORMANCE OF COMMUNITY STRUCTURE DETECTION

Number of nodes	Modularity Q of detected community structure		
	h=1	*2*	*3*
200	0.2419	0.2914	0.2010
400	0.2760	0.2502	0.1749
600	0.2406	0.2195	0.1403
800	0.2108	0.1870	0.1171
1000	0.1913	0.1482	0.0864
1200	0.1500	0.1207	0.0801

[1] http://www-personal.umich.edu/~mejn/netdata/

The experiment result shows that optimal DBN structure of power dataset is 1000 input nodes, 2 hidden layers which conclude 400 and 200 nodes respectively and 100 classification layer nodes.

Fig. 4 and Fig. 5 show the time consuming of GN and SymNMF on dataset blogcatalog with 10312 nodes, and sparsity of the dataset is 0.014%. We extract first K nodes as representative set to detect the community structure via GN and SymNMF respectively. The configuration of experimental computer is Intel I7 processer, 8G memory and Win7 64bit OS. The time consuming of GN and SymNMF algorithm increase rapidly and cannot detect large-scale network. The reason is that GN traverses different number of communities and calculates modularity of community. Besides, SymNMF expect to set the number of community manually.

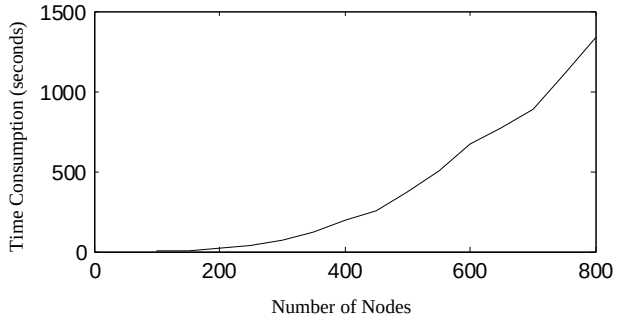

Fig.4. Efficiency of GN algorithm

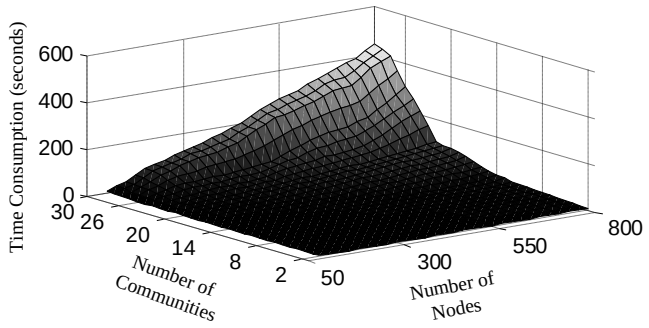

Fig.5. Efficiency of SymNMF algorithm

The advantages of the proposed algorithm are indicated in Table 3. The feature dimension is compressed by Bernoulli random measurement matrix. Enabling DBN by setting it to 1000 input nodes, 2 hidden layers are respectively 400 and 200 and 100 classification layer nodes. As depicted in Table 3, for three real-world data sets, it spent the majority of time on detect community structure of representative set by GN. However, because of control the size of representative set, the algorithm running time do not change radically as the scale increases. With the increasing scale of network, the time of model training and recognition increases linearly. For example, we can still finish training and recognition tasks in the effective time when the scale is 80513. It is proved that the proposed method overcomes the limit of handling large-scale dataset.

TABLE III. EFFICIENCY OF PROPOSED ALGORITHM UNDER REAL-WORLD DATASETS

Name of Real-world datasets	Scale of datasets (Number of nodes)	Time consumptions (seconds)		Modularity Q	
		Representative sets	Universal sets	Representative sets	Universal sets
power	4914	1915.1	116.5	0.7233	0.2914
blogcatalog[2]	10312	1888.2	226.3	0.6874	0.2328
flickr[3]	80513	1861.9	1652.9	0.7025	0.2077

IV. Conclusion

An effective community structure detection method was proposed, and its power verified by experiments. Compared to conventional community detection algorithms, the proposed algorithm produces improved performance of handling large-scale social network data. The experiment results encourage the use of the proposed algorithm in many practical applications. The intention for future work is to extract community feature from unsupervised training results directly. This can avoid incomprehensiveness of the representative set which lead to insufficient supervised training.

REFERENCES

[1] Girvan M, Newman M E J. Community Structure in Social and Biological Networks[J]. The National Academy of Sciences, USA, 2002, 99(12): 7821-7826.

[2] Newman M E J. Fast Algorithm for Detecting Community Structure in Networks[J]. The American Physical Society, 2004, 69(6): 066133.

[3] Kuang D, Ding C, Park H. Symmetric Nonnegative Matrix Factorization for Graph Clustering[C]//Proceedings of 2012 SIAM International Conference on Data Mining. Anaheim, USA, 2012: 106–117.

[4] Shiga M, Takigawa I, Mamitsuka H. A Spectral Clustering Approach to Optimally Combining Numerical Vectors with a Modular Network[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2007:647-656.

[5] Donoho D L. Compressed Sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[6] Candes E J. Compressive Sampling[C]//Proceedings of the International Congress of Mathematicians. Madrid, Spain, 2006:1434-1452.

[7] Hinton G, Osindero S. A fast learning algorithm for deep belief nets [J].Neural Computation, 2006, 18(7):1527-1554.

[8] Quoc V. L, Marc A R, Rajat M, et, al. Building High-level Features Using Large Scale Unsupervised Learning[C]. The 29'th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.

[9] Mohamed A, Hinton G, Penn G. Understanding How Deep Belief Network Perform Acoustic Modelling[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, 2012: 4273-4276.

[10] Candes E J. The Restricted Isometry Property and Its Implications for Compressed Sensing[J]. Computes Rendus Mathematique, 2008, 346(9-10): 589-592.

[11] Baraniuk R, Davenport M, Devore R, Wakin M. A Simple Proof of the Restricted Isometry Property for Random Matrices[J]. Constructive Approximation, 2008.12, 28(3): 253-263.

[12] Newman M E J, Girvan M. Finding and Evaluating Community Structure in Networks[J]. Physical Review E, 2004, 69(2): 026113.

[2] http://socialcomputing.asu.edu/datasets/BlogCatalog3

[3] http://socialcomputing.asu.edu/datasets/Flickr