Behavior-Based Collective Classification in Sparsely Labeled Networks

Gajula.Ramakrishna[1], M.Kusuma[2]
[1]*Gajula.Ramakrishna, Dept of Master of Computer Applications, EAIMS, Tirupathi, India*
[2]*Professor, Dept of Master of Computer Applications, EAIMS, Tirupathi, India*

***Abstract*- Classification in sparsely labeled networks is challenging to traditional neighborhood-based methods due to the lack of labeled neighbors. In this paper, we propose a novel behavior-based collective classification (BCC) method to improve the classification performance in sparsely labeled networks. In BCC, nodes' behavior features are extracted and used to build latent relationships between labeled nodes and unknown ones. Since mining the latent links does not rely on the direct connection of nodes, decrease of labeled neighbors will have minor effect on classification results. In addition, the BCC method can also be applied to the analysis of networks with heterophily as the homophily assumption is no longer required. Experiments on various public data sets reveal that the proposed method can obtain competing performance in comparison with the other state-of-the-art methods either when the network is labeled sparsely or when homophily is low in the network.**

***Index Terms*- Behavior feature, sparsely labeled networks, collective classification, within-network classification.**

I. INTRODUCTION

Given a partially labeled network, in which labels of some nodes are known, within-network classification aims to predict labels of the rest nodes. Due to the increasingly wide applications in counter terrorism analysis[1],[2],fraud detection [3], [4] and product recommendations [5], [6] etc., within-network classification has received a lot of attention in recent years. Conventional classification methods assume the data is independent and identically distributed (i.i.d.). Nevertheless, in network data ,then connected with each other, making the label of nodes are correlated with not only its own attributes, but also the label of neighbors [7]–[11]. For example, wvRN [7], [8] predicts the label of unknown nodes via a weighted average of the estimated class membership of the node's neighbors. In a range of real networks, wvRN has shown to obtain a surprisingly good performance [7]. How ever ,wvRNrelies heavilyon the homophily assumption, i.e., nodes belonging to the same class tend to be linked with eachother [12],and there by are limited in the analysis of networks where nodes are not clustered by the studied property. Probabilistic relational models [9]–[11] can overcome thisLimitation. In probabilistic relational models, by constructing the dependence between connected nodes, the probability of an unknown node's label is conditioned not only on the labels of its neighbor nodes, but also on all observed data (i.e., network structure and all labeled nodes).

II. RELATED WORK

SEMI-SUPERVISED LEARNING

Making use of both labeled and unlabeled data, semi-supervised learning is an effective method for classification in sparsely labeled networks [22], [23]. One type of this method is to design a classification function which is sufficiently smooth with respect to the intrinsic structure collectively revealed by labeled and unlabeled points [24]. Zhou *et al*. [24] propose a simple iteration algorithm, which considered global and local consistency by introducing a regularization parame- ter. By modeling the network with constraint on label consis- tency, Zhu *et al*. [25] propose a Gaussian random field (GRF) method by introducing a harmonic function, of which the value is the average of neighboring points. Another type of semi-supervised learning methods is the graph-cut method [26]–[28], which assumes that more closely con- nected nodes tend to belong to the same category. The core idea is to find a cut set with the minimum weight by using different criteria. However, the high cost of computing often lead to poor performance of the algorithm when applied in large networks. Some other algorithms use random walk on the network to obtain a simple and effective

solution by propagating labels from labeled nodes to unknown nodes. Based on passaging time during random walks with bounded lengths, Callut *et al*. [29] and Newman [30] introduce a novel technique, called D-walks, to handle semi-supervised classi- fication problems in large graphs. Zhou and Schlkopf [31] define calculus on graphs by using spectral graph theory, and propose a regularization framework for classification problems on graphs. However, many semi-supervised learn- ing methods rely heavily on the assumption that the net- work exhibits homophily, i.e., nodes belonging to the same class tend to be linked with each other [12]. Meanwhile, the implementation of semi-supervised learning algorithm often requires a large amount of matrix computation, and thus is infeasible for processing large datasets [25]. Many methods have been developed to overcome these limitations. For example, Tong *et al*. propose a fast random walk with restart algorithm [32] to improve the performance on large- scale dataset. Lin *et al*. propose a highly scalable method, called Multi-Rank-Walk (MRW), which requires only linear computation time in accordance to the number of edges in the network [12]. Mantrach *et al*. [33] design two iterative algorithms which can be applied in networks with millions of nodes to avoid the computation of the pairwise similarities between nodes. Gallagher *et al*. [13] design an even-step random walk with restart (Even-step RWR) algorithm, which mitigates the dependence on network homophily effectively.

A. ACTIVE LEARNING

In active learning [34], the number of known labels required for accurate learning is reduced by intelligently selecting to- be-labeled nodes to achieve improved classification perfor- mance in sparsely labeled networks. Lewis and Catlett [35] propose a method based on uncertainty reduction, which selects the data with lowest certainty for querying. However, the method will fail when there are a certain number of outliers. The outliers have high uncertainty in the network, but getting their labels doesn't help to inference the rest data. To handle this limitation, Roy and McCallum [36] design a method to determine the impact on the expected error of each potential labeling request by using Monte Carlo approach. In the active learning process, the feature of linked data in the network can also be taken into account. Bilgic and Getoor [18] propose several ways of adapting existing active learning methods to network data. Macskassy [37] designs a novel hybrid approach by using community detection and social network analytic centrality measures to identify the candidates for labeling. When network structure and node attribute information are available, Bilgic *et al*. [19] apply several classic active learning strategies such as disagree- ment and clustering to select samples for labeling, which has shown significant improvements over baseline methods. Active learning is able to overcome the sparse labeling prob- lem to some extent, but it still requires the participation of experts and lacks an automatic learning process.

III. METHOD

In this section, we will describe the intuition of behavior based classification at first ,and show that the behavior feature is more discriminative than traditional similarity measures. Then, the framework of our method is introduced in detail.

A. INTUITION

In sparsely labeled networks, the labels of nodes are much fewer, making it difficult to leverage label dependencies to make accurate prediction. Without considering the label information, it can be found that the network structure can still provide useful information. Therefore, most researches focus on utilizing the network structure to predict unknown nodes. For example, CN method [21] estimates the similarity of nodes by local structure (the number of common neighbors). However, it becomes ineffective when handling the sparsely labeled network classification task in some situations. Figure 1 shows a sparsely labeled network, in which only node a and node b are labeled and the task is to predict the label of no deu (the true coloris‘‘red’’).C N method considers that node u has two common neighbors with node a, so the similarity between node u and node a is 2. Then we can find that the similarity between node u and node b is 2 as well. In this situation, CN method cannot determine which is the most similar node with u, and thus, leading to lower performance.

B. BEHAVIOR BASED COLLECTIVE CLASSIFICATION

Since behavior feature can provide a different kind of information that may be useful in sparsely labeled networks, we propose a novel Behavior-based Collective Classification method (BCC) in this paper to handle the sparse labeling problem. The process of BCC in network data consists of four steps: behavior feature extraction, screening valuable nodes, classification by voting and collective inference.

IV. IMPLEMENT

BCC method consists of four steps for classification, and in this section, we introduce the implement of each step in detail. Firstly, we will describe how to extract behavior feature, which has shown more discriminative ability in sparsely labeled networks. In order to handle the imbalanced dataset, we only allow the most relevant nodes in the classification process by using correlation and similarity analysis. Then we introduce the strategy of voting for classification. Collective inference procedure is used to handle the extremely sparse labeling problem, which is described afterwards. Finally, the algorithm is given to show the details of our method.

A. BEHAVIOR FEATURE EXTRACTION

Let w(i,j) be the weight of the edge from node it on then the adjacency vector can be used to describe the behavior pattern of node i. However, it should be noted that E wi is the observed value in the current time, which may change by time with the evolution of network. Therefore ,instead of using E wi ,we need to extract more stable behavior feature to be able to reflect nodes' intrinsic attribute.

B. SCREEN VALUABLE NODES FOR CLASSIFICATION

The labeled nodes are much fewer in sparsely labeled network, so traditional methods tend to utilize all the labeled nodes in the classification process. However, involving unrelated nodes in the classification process will only bring noise data and lead top or performance. Moreover, when classes of labeled nodes are imbalanced, unknown nodes will be more likely to be labeled the same as the majority. To solve this issue, we show how to find them relevant nodes, from the perspective of correlation and similarity of behavior feature, to reduce the impact of noise code.

V. EXPERIMENTAL SETUP

DATASETS:

We evaluate the proposed BCC method by comparing it with other baseline methods for the classification performance on the following real-world datasets. 1, Enron emails [46]. We choose 151 persons as nodes in the network and retain 2235 edges connected between these persons. 102 out of these 151 nodes are assigned a role according to the role list [47], where 37 nodes are labeled as ''employee''. The network is a small directed weighted graph composed of email communication among users and the experimental task is to identify the ''employee'' class. 2, Web KB [48]. Web KB is a dataset of web pages gathered from different universities, in which nodes are web pages and edges are hyperlinks. Each webpage is classified into one of the five classes: ''course, faculty, student, project, staff''. There are four networks in this dataset: cornell, tex as, Washington and wisconsin, and the task is to identify the''student'' webpage. 3,Cora[48].Corais acitation network forme dof2708scientific publications and 5429 links. Each publication is classified into one of seven classes. The task is to identify the ''Neural_Networks'' class. 4, Citeseer [48]. Citeseer is a citation network consists of 3312 scientific publications and 4732 links. Each publication is classified in to one of six classes. The task is to identify the ''DB'' class.

VI.CONCLUSION

In order to improve classification accuracy in sparsely labeled networks ,we propose a novel behavior based collective classification method, BCC, in this study. In BCC, the behavior feature of nodes is extracted for classification, which has shown more discriminative ability to traditional methods. Then, instead of using all the labeled nodes, we screen the most-relevant nodes according to the calculation of correlation and similarity, which can overcome the effects of noise and imbalanced dataset. Finally, collective inference is introduced to utilize both labeled nodes and unlabeled nodes, which can relieve the sparse labeling problem effectively. Extensive experiments on public data set demonstrate that BCC method outperforms several baseline methods, especially when the network is

sparsely labeled. Meanwhile, instead of relying on local neighbor nodes, BCC method predicts unknown nodes by using valuable nodes which may not even connected directly, making it a preferable method for classification in networks with heterophily. Note that in Enron dataset, only a subset of nodes have labels and we can only compare different methods on these nodes ,but unlabeled nodes and their connections to labeled nodes may still provide useful behavior information, which can be utilized in BCC method. From this point of view, BCC shares the similar idea with semi-supervised learning. The current implementation of BCC has limited computing efficiency for similarity comparison, when the network is large, it may become a bottleneck for the algorithm. Future work may also model the network with different generation process, and other types of behavior feature and strategies in the classification process may be applied. Another challenging extension is the multi-label classification in sparsely labeled networks, where instances can be assigned with multiple labels and the labeled nodes are few in the network. We believe this study highlights the importance of behavior feature in improving performance of network classification and the BCC method could be used in a variety of settings with generalized stability.

REFERENCES

[1] S. A. Macskassy and F. Provost, ‘‘A brief survey of machine learning methods for classification in networked data and an application to suspicion scoring,’’ in Statistical Network Analysis: Models, Issues, and New Directions. Berlin, Germany: Springer, 2007, pp. 172–175.

[2] S. A. Macskassy and F. Provost, ‘‘Suspicion scoring based on guilt-byassociation, collective inference, and focused data access,’’ in Proc. Int. Conf. Intell. Anal., 2005

[3] S. Hill, D. K. Agarwal, R. Bell, and C. Volinsky, ‘‘Building an effective representation for dynamic networks,’’ J. Comput. Graph. Statist., vol. 15, no. 3, pp. 584–608, 2006.

[4] J. Neville, Ö. Şimşek, D. Jensen, J. Komoroske, K. Palmer, and H. Goldberg, ‘‘Using relational knowledge discovery to prevent securities fraud,’’ in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2005, pp. 449–458.

[5] P. Domingos and M. Richardson, ‘‘Mining the network value of customers,’’ in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2001, pp. 57–66.

[6] Z. Huang, H. Chen, and D. Zeng, ‘‘Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering,’’ ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 116–142, 2004.

[7] S. A. Macskassy and F. Provost, ‘‘Classification in networked data: A toolkit and a univariate case study,’’ J. Mach. Learn. Res., vol. 8, pp. 935–983, May 2007.

[8] S. A. Macskassy and F. Provost, ‘‘A simple relational classifier,’’ in Proc. 2nd Workshop Multi-Relational Data Mining (MRDM) , 2003, pp. 1–13.

[9] B.Taskar,E.Segal,andD.Koller,‘‘Probabilisticclas sificationandclustering in relational data,’’ in Proc. Int. Joint Conf. Artif. Intell., vol. 17. 2001, pp. 870–878.

[10] B. Taskar, P. Abbeel, and D. Koller, ‘‘Discriminative probabilistic models for relational data,’’ in Proc. 18th Conf. Uncertainty Artif. Intell., 2002, pp. 485–492.

[11] J. Neville and D. Jensen, ‘‘Collective classification with relational dependency networks,’’ in Proc. 2nd Int. Workshop Multi-Relational Data Mining, 2003, pp. 77–91.

[12] F. Lin and W. W. Cohen, ‘‘Semi-supervised classification of network data using very few labels,’’ in Proc. Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2010, pp. 192–199.

[13] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, ‘‘Using ghost edges for classification in sparsely labeled networks,’’ in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 256–264.

[14] R. Xiang and J. Neville, ‘‘Pseudolikelihood em for within-network relational learning,’’ in Proc. 8th IEEE Int. Conf. Data Mining (ICDM), Dec. 2008, pp. 1103–1108.

[15] J.J.Pfeiffer,J.Neville,andP.N.Bennett,‘‘Overcomi ngrelationallearning biases to accurately predict preferences in large scale networks,’’ in Proc.

24th Int. Conf. World Wide Web, 2015, pp. 853–863.

[16] L. K. Mcdowell, ‘‘Relational active learning for link-based classification,’’ in Proc. IEEE Int. Conf. Data Sci. Adv. Anal., Oct. 2015, pp. 1–10.

[17] A. Kuwadekar and J. Neville, ‘‘Relational active learning for joint collective classification models,’’ in Proc. Int. Conf. Mach. Learn. (ICML), Bellevue, WA, USA, Jun./Jul. 2012, pp. 385–392.

[18] M. Bilgic and L. Getoor, ‘‘Link-based active learning,’’ in Proc. NIPS Workshop Anal. Netw. Learn. Graph., 2009, pp. 1–7.

[19] M. Bilgic, L. Mihalkova, and L. Getoor, ‘‘Active learning for networked data,’’ in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010, pp. 79–86.

[20] S. A. Macskassy, ‘‘Improving learning in networked data by combining explicit and mined links,’’ in Proc. Nat. Conf. Artif. Intell., vol. 22. Menlo Park, CA, USA, 2007, p. 590.