Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-attribute and relational learning via hypergraph regularized generative model

Shaokai Wang [a], Xutao Li [a], Yunming Ye [a,*], Xiaohui Huang [b], Yan Li [c]

[a] Department of Computer Science, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[b] School of Information Engineering Department, East China Jiaotong University, Nanchang, China
[c] School of Computer Engineering, Shenzhen Polytechnic, Shenzhen, China

ARTICLE INFO

Article history:
Received 29 February 2016
Received in revised form
27 May 2016
Accepted 3 June 2016

Keywords:
Hypergraph learning
Collective classification
Multiple attributes
Semi-supervised learning

ABSTRACT

The real-world networking data may contain different types of attribute views and relational view. Hence, it is desirable to collectively use available attribute views and relational view in order to build effective learning models. We call this framework multi-attribute and relational learning. Collective classification is one of the popular approaches that can handle both attribute and relational information for network data. However, in collective classification only one type of attribute and relational view is involved and little attention is received for multi-attribute and relational learning. In this paper, we propose a new semi-supervised collective classification approach, called hypergraph regularized generative model (HRGM), for multi-attribute and relational learning. In the approach, a generative model based on the Probabilistic Latent Semantic Analysis (PLSA) method is developed to leverage attribute information, and a hypergraph regularizer is incorporated to effectively exploit higher-order relational information among the data samples. Experimental results on various data sets have demonstrated the effectiveness of the proposed HRGM, and revealed that our approach outperforms existing collective classification methods and multi-view classification methods in terms of accuracy.

1. Introduction

Collective classification is an important research problem in data mining, which aims to predict the labels of nodes in terms of their attribute and network topology. In the past decade, many methods have been developed to deal with the problem [1–6]. These methods integrate the attribute and topology information to build classifiers for prediction.

Previous collective classification approaches usually assume that network data has only one type of node attribute and one type of relation [1,2,5,6]. However, this is clearly a strong limitation in many real-world applications, where complex network structures (e.g. multiple types of attributes, multiple types of relations) are available. For example, in Twitter system, to annotate interests for users, they can be represented by multiple attribute views such as their tweets, hashtags and list memberships; at the same time, different Twitter users can be connected with follower, followed by and re-tweet links. Hence, network data can have multiple attributes and multiple relational representations. We refer to networks of such type as multi-attribute network data. Although two methods are developed to tackle collective classification for multi-attribute network data in recent studies [3,4], both require conditional independent assumption of attribute views to work well, which may be too strong to be satisfied in real applications. In this paper, we aim to solve the collective classification problem for multi-attribute network.

For the collective classification problem in a multi-attribute network, there are three main challenges: (1) how to utilize the multi-view attributes; (2) how to exploit the relational network topology; (3) preparing the label information is often costly and thus learning with sparsely labeled data is a desire. A good multi-attribute and relational learning model should address the three challenges simultaneously.

Recently, many researchers address the three challenges separately. *Challenge* 1: many learning techniques have been proposed to utilize multi-view attribute for classification [7–10]. The core idea of these techniques is integrating complementary information in different views for better classification performance. Co-training [8] is one of the pioneering schemes in this field. It trains two classifiers alternatively to maximize the mutual agreement on two distinct views of

* Corresponding author.
E-mail address: yeyunming@hit.edu.cn (Y. Ye).

http://dx.doi.org/10.1016/j.neucom.2016.06.003
0925-2312/

unlabeled data. Recently, Zhuang et al. [9] proposed MVPLSA algorithm, which is a multi-view extension to Probabilistic Latent Semantic Analysis. The MVPLSA algorithm jointly models the co-occurrences of features and documents. However, these multi-view learning approaches are built merely on the basis of attributes, and not applicable to network data. *Challenge* 2: graph-based approaches gain great popularity in exploiting network topology for classification [5,6,11,12]. These models are on the basis of linked instances having similar labels, and they make use of regularizer to constrain label distributions of two linked instances. However, such pairwise constraints may ignore the higher-order statistical information among instances. Recent studies show that modeling higher-order information will improve the classification performance [13–15]. *Challenge* 3: when the labeled data is sparse, a wise solution is adopting semi-supervised learning scheme to exploit unlabeled data. Although some collective classification methods are developed for a semi-supervised purpose [5,6,16–19], they consider single view feature and cannot deal with multi-attribute network data, and also higher-order information among instances is ignored.

In this paper, we propose a new approach called HRGM for collective classification in a sparsely labeled multi-attribute network. The approach fuses multi-view attributes and higher-order relational information by a *H*ypergraph *R*egularized *G*enerative *M*odel (HRGM). Specifically, our method utilizes multi-view features by a generative model. To exploit the network topology and unlabeled data, we incorporate a hypergraph regularizer into the generative model. On the one hand, the hypergraph regularizer is able to model higher-order information among linked instances. On the other hand, it can make use of unlabeled data to propagate label information by means of semi-supervised learning.

Our main contribution in this paper can be summarized as follows: (i) We propose a new learning algorithm (HRGM) for solving the problem of collective classification in a multi-attribute network. This model is capable of utilizing both the multi-view features and higher-order topology information; meanwhile the unlabeled data can be exploited for prediction. (ii) An effective EM algorithm is derived to compute the solution of our HRGM model. (iii) The classification performance of the proposed algorithm is verified through comprehensive experiments, and the results show that the proposed HRGM is effective to fuse attribute and higher-order relational information and delivers promising classification performance.

The paper is organized as follows. In Section 2, we review the related work. In Section 3, we present our proposed HRGM method. Experimental results are given in Section 4. We conclude this paper in Section 5.

2. Related work

2.1. Semi-supervised collective classification

In the last decade, collective classification has received considerable attention. Most of the collective classification approaches assume that sufficient labeled data are available. However, obtaining sufficient labeled data is expensive, or even not possible in many cases [17]. To address this issue, some recent studies adopt semi-supervised learning to leverage the unlabeled portion of a sparsely-labeled network for classification [5,6,16–19]. McDowell and Aha [17] argued that semi-supervised learning could improve the performance of iterative learning methods. They proposed a hybrid model that uses one classifier for node features and another classifier for relational features. They demonstrated that the hybrid model with label regularization can significantly increase prediction accuracy. Shi et al. [5] proposed a probabilistic generative model with network regularization for semi-supervised collective classification. This model computes label probability distributions for unlabeled instances, by taking network structures into account. Wu et al. [6] proposed a multi-label regularized generative model for semi-supervised collective classification. They show that the leveraging unlabeled data is essential to achieving high performance for semi-supervised collective classification. However, these approaches are designed for single attribute network data, and cannot be directly used for multi-attribute network data. In addition, these approaches ignore the higher-order correlations among instances.

There are only a few methods designed for multi-attribute network data [3,4]. Shi et al. [3] proposed a gradient boosting consensus (GBC) approach for multi-attribute network classification. GBC takes advantage of vector-based features and graph structure features simultaneously. It generates decision trees collectively from all data sources in each iteration, and all of the trees are aggregated together to make the final prediction. Vijayan et al. [4] proposed a unified multi-label model using multiple attribute views on multi-relational network data, which captures complex label correlations within and across attribute/relationship types. The algorithm tries to maximize the consensus among various attribute and relational views, and simultaneously reduces disagreement between attribute and relational views. The two methods [3,4] are all co-training style algorithms, which requires the conditional independent assumption to work well, and the assumption may not be true in some real applications. Moreover, neither of the approaches can deal with higher-order relationships among instances.

2.2. Multi-view learning

Multi-view learning is a hot topic and many algorithms have been proposed [8,9,20–22]. Blum and Tom [8] initially proposed the idea of co-training to solve the semi-supervised learning with two view features. It trains two classifiers alternatively to maximize the mutual agreement on the two distinct views of the unlabeled data. Following this idea, many variants have already been developed [20–22]. Kumar et al. [20] developed a multi-view spectral clustering algorithm utilizing the idea of co-training. Nie et al. [23] proposed an approach to model the progression of chronic diseases based on multimedia and multimodal observational data. Akbari et al. [24] proposed an adaptive multi-modal reranking model, which is able to jointly regularize the relatedness among different modalities. Song et al. [25] proposed a structure-constrained multi-source multi-task learning scheme in the context of user interest inference. Song et al. [26] developed a model for missing value completion in multiple social networks, and utilized the completion results for volunteerism tendency prediction.

As a representative work, Zhuang et al. [9] proposed MVPLSA, a multi-view learning algorithm based on Probabilistic Latent Semantic Analysis. The MVPLSA algorithm is motivated by the following two observations. First, some features may be grouped together to form a high-level topic (feature clusters), e.g., the words “price”, “performance”, “announcement” from an enterprise news may present the concept “product announcement”; second, the methods working with only one single view may not perform well, since they cannot make full use of the knowledge from other views. The MVPLSA jointly models the co-occurrences of features and documents. In the model, there are two latent variables, y for the high-level latent topic and z for the document cluster, and three visible variables, d for the document, f for the feature, and v for the view. The conditional probability $p(z|d)$, which is

Please cite this article as: S. Wang, et al., Multi-attribute and relational learning via hypergraph regularized generative model, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.06.003