Neurocomputing 000 (2017) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised multi-graph classification using optimal feature selection and extreme learning machine

Jun Pang[a,b,c,*], Yu Gu[c], Jia Xu[d], Ge Yu[c]

[a] College of Computer Science and Technology, Wuhan University of Science and Technology, Hubei 430065, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Hubei, 430065, China
[c] College of Computer Science and Engineering, Northeastern University, Liaoning 110169, China
[d] School of Computer, Electronics and Information, Guangxi University, Guangxi 530004, China

ARTICLE INFO

Article history:
Received 30 September 2016
Revised 27 December 2016
Accepted 4 January 2017
Available online xxx

Keywords:
Multi-graph
Semi-supervised
Feature selection
Extreme learning machine

ABSTRACT

A multi-graph is represented by a bag of graphs. Semi-supervised multi-graph classification is a partly supervised learning problem, which has a wide range of applications, such as bio-pharmaceutical activity tests, scientific publication categorization and online product recommendation. However, to the best of our knowledge, few research works have be reported. In this paper, we propose a semi-supervised multi-graph classification algorithm to handle the semi-supervised multi-graph classification problem. Our algorithm consists of three main steps, including the optimal subgraph feature selection, the subgraph feature representation of multi-graph and the semi-supervised classifier building. We first propose an evaluation criterion of the optimal subgraph features, which not only considers unlabeled multi-graphs but also considers the constraints between the multi-graph level and the graph level. Then, the optimal subgraph feature selection problem is equivalently converted into the problem of mining m most informative subgraph features. Based on those derived m subgraph features, every multi-graph is represented by an m-dimensional vector, where the ith dimension equals to 1 if at least one graph involved in the multi-graph contains the ith subgraph feature. At last, based on these vectors, semi-supervised extreme learning machine(semi-supervised ELM) is adopted to build the prediction model for predicting the labels of unseen multi-graphs. Extensive experiments on real-world and synthetic graph datasets show that the proposed algorithm is effective and efficient.

1. Introduction

A multi-graph is modeled as a bag of graphs. It is a powerful model to represent complicated structures of objects in physical world. For example, a paper can be represented as a multi-graph, shown as in Fig. 1. A labeled multi-graph is a multi-graph with a class label. If once one of graphs in a multi-graph is labeled as positive, the multi-graph is labeled as positive. Otherwise, the multi-graph is labeled as negative. Multi-graph classification problem aims to learn a prediction model with the aid of labeled multi-graphs and to predict the class labels of those unlabeled multi-graphs, having many practical application, including drug activity detection [1,2], science publication classification [1,2] and product recommendation [1,2]. Two application examples of the multi-graph classification are shown in Example 1 and Example 2, respectively.

Example 1. Given a collection of papers, we use the domain fields of papers, such as Artificial Intelligence, Computer Vision and so on, as the class labels of these papers. The multi-graph classification methods can be used to predict the domain fields of unlabeled papers with these labeled papers [1,2].

Example 2. A molecule has many forms. If one of its forms resists a certain disease, the molecule in this form can be used to manufacture drugs to cure such disease. The specific form of a molecule can be described as a graph. Under such circumstances, a multi-graph represents different forms of the molecule. The multi-graph classification algorithms can be utilized to predict the molecules activities [1,2].

However, because data often are labeled through manual works which are time-consuming with high-cost, it is usually unrealistic to get many labeled multi-graphs in practice, which leads to low classification accuracy. Although it is difficult to get labeled multi-

* Corresponding author at: College of Computer Science and Technology, Wuhan University of Science and Technology, Hubei 430065, China.
E-mail addresses: pangjun@wust.edu.cn (J. Pang), guyu@mail.neu.edu.cn (Y. Gu), xujia@gxu.edu.cn (J. Xu), yuge@mail.neu.edu.cn (G. Yu).

http://dx.doi.org/10.1016/j.neucom.2017.01.114
0925-2312/

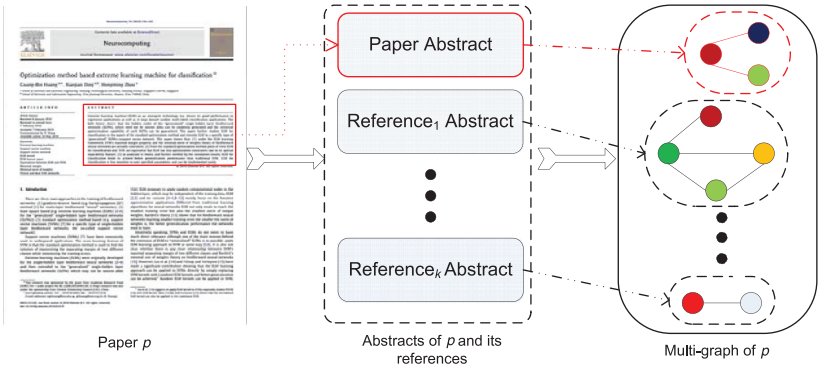

Fig. 1. An example of multi-graph representation model. (Paper p is modeled as a multi-graph, where each graph inside the multi-graph corresponds to the abstract of p or the abstract of the reference cited by p. A graph is constructed by using keywords of the abstract as nodes and their correlations as edges.)

graphs, it is much easier to collect unlabeled multi-graphs. It may be an alternative solution to train the classifier not only by the limited labeled multi-graphs but also by the rich unlabeled multi-graphs. In other words, the semi-supervised learning methods can be used here to improve the classification accuracy of the multi-graph classification problem.

The target of multi-graph semi-supervised learning is to learn a classifier from a small amount of labeled multi-graphs and a large amount of unlabeled multi-graphs to predict the class labels of unlabeled multi-graphs as accurately as possible. Inspired by the methods of semi-supervised subgraph feature selection [3] and multi-graph classification [1,2], we first obtain optimal subgraph features from the multi-graph dataset based on our criterion. Optimal subgraph features are subgraphs, which are most useful to distinguish multi-graphs. Then, every multi-graph is represented by an m-dimensional vector, where the jth dimension equals to 1 if at least one graph involved in the multi-graph contains the jth subgraph feature. Lastly, based on these vectors, an existing semi-supervised algorithm is utilized to build a classifier. After unseen multi-graphs are represented by vectors based on these selected optimal subgraph features, their labels are predicted using the built classifier. It is a non-trivial task to carry out semi-supervised multi-graph classification, mainly facing the following two challenges.

1. Multi-graph partially supervised learning can be divided into two categories [4] i.e., positive and unlabeled learning, and labeled and unlabeled learning (semi-supervised learning). In positive and unlabeled learning, the class labels of all labeled multi-graphs are positive. While, in the second category, the class labels of labeled multi-graphs are not only involve positive labels but also involve other labels. To the best of our knowledge, hardly any works about semi-supervised multi-graph classification have been reported. Wu et al. [5] propose a new learning framework puMGL to solve positive and unlabeled multi-graph classification problem, which obtains higher classification accuracy than the baseline methods. However, puMGL is specially designed to process the datasets which only contain a small number of positive multi-graphs.
2. In order to gain more profits, many practical applications, such as marketing decision and product recommendation, require that the prediction algorithms can ensure a high classification accuracy. However, using existing subgraph feature evaluation criterions [1–3] and the traditional semi-supervised classification models [4] can only achieve a low classification accuracy. Because existing solutions need a large amount of labeled multi-graphs. However, in our problem setting, there are only a small amount of labeled multi-graphs. In addition, existing subgraph feature evaluation criterions for semi-supervised graph classification problem do not consider the constrains of the multi-graph level.

Facing these challenges, we propose an algorithm to solve the semi-supervised multi-graph classification problem. In order to improve the classification accuracy, we propose an novel evaluation criterion to select optimal subgraph features, and adopt semi-supervised ELM to build the prediction model. Compared to the subgraph feature evaluation criterions of supervised multi-graph classification [1,2], our evaluation criterion not only considers labeled multi-graphs but also considers unlabeled multi-graphs. Compared to the subgraph feature evaluation criterion of semi-supervised graph classification [3], our evaluation criterion considers both of the constrains of the graph level and the constrains of the multi-graph level. Moreover, since semi-supervised ELM has a good performance [6]. It is adopted to further improve the classification accuracy in this paper. In addition, a upper bound strategy is derived to improve the efficiency. The major contributions of this paper are summarized as follows.

1. We propose a evaluation criterion of optimal subgraph feature, based on which we design an algorithm to select the optimal subgraph features. In addition, an upper bound pruning strategy is proposed to improve the efficiency of feature selection.
2. We adopt semi-supervised ELM to improve the classification accuracy. Moreover, we propose an algorithm based on our subgraph feature selection algorithm and semi-supervised ELM to solve the semi-supervised multi-graph classification problem.
3. We have conducted extensive experiments on both real and simulated data sets to verify the effectiveness and efficiency of our proposals.

The remainder of this paper is organized as follows. Related works are introduced in Section 2. Problem definitions are discussed in Section 3. The proposed algorithms are provided in Section 4. Experimental results and discussions are presented in Section 5. We conclude the paper in Section 6.

2. Related works

Related works of our study include multi-graph partially supervised learning, subgraph feature selection and semi-supervised extreme learning machine.

2.1. Multi-graph partially supervised learning

Some works about positive and unlabeled multi-graph classification have been reported. Wu et al. [5] propose a learning framework, called puMGL, to solve positive and unlabeled multi-graph classification problem. A subgraph feature selection metric is designed to gain higher accuracy. Inspired by traditional positive and

Please cite this article as: J. Pang et al., Semi-supervised multi-graph classification using optimal feature selection and extreme learning machine, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.01.114