Pattern Recognition 48 (2015) 3783–3796

Contents lists available at ScienceDirect

Pattern Recognition

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

Finding the best not the most: regularized loss minimization subgraph selection for graph classification

Shirui Pan [a,*], Jia Wu [a], Xingquan Zhu [b], Guodong Long [a], Chengqi Zhang [a]

[a] Centre for Quantum Computation & Intelligent Systems, FEIT, University of Technology Sydney, Australia
[b] Department of Computer and Electrical Engineering & Computer Science, Florida Atlantic University, USA

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form
15 May 2015
Accepted 18 May 2015
Available online 31 May 2015

Keywords:
Feature selection
Classification
Graph classification
Sparse learning

ABSTRACT

Classification on structure data, such as graphs, has drawn wide interest in recent years. Due to the lack of explicit features to represent graphs for training classification models, extensive studies have been focused on extracting the most discriminative subgraphs features from the training graph dataset to transfer graphs into vector data. However, such filter-based methods suffer from two major disadvantages: (1) the subgraph feature selection is separated from the model learning process, so the selected most discriminative subgraphs may not best fit the subsequent learning model, resulting in deteriorated classification results; (2) all these methods rely on users to specify the number of subgraph features K, and suboptimally specified K values often result in significantly reduced classification accuracy.

In this paper, we propose a new graph classification paradigm which overcomes the above disadvantages by formulating subgraph feature selection as learning a K-dimensional feature space from an *implicit* and *large* subgraph space, with the optimal K value being automatically determined. To achieve the goal, we propose a regularized loss minimization-driven (RLMD) feature selection method for graph classification. RLMD integrates subgraph selection and model learning into a unified framework to find discriminative subgraphs with guaranteed minimum loss *w.r.t.* the objective function. To automatically determine the optimal number of subgraphs K from the exponentially large subgraph space, an effective *elastic net* and a subgradient method are proposed to derive the stopping criterion, so that K can be automatically obtained once RLMD converges. The proposed RLMD method enjoys gratifying property including proved convergence and applicability to various loss functions. Experimental results on real-life graph datasets demonstrate significant performance gain.

1. Introduction

Recent years have witnessed an increasing number of applications involving objects with structural relationships, including chemical compounds in Bioinformatics [1], brain networks [2], image structures [3], and academic citation networks [4]. For these applications, graph is a natural and powerful tool for modeling and capturing dependency relationships between objects.

Unlike conventional data, where each instance is represented in a feature-value vector format, graphs exhibit node–edge structural relationships and have no natural vector representation[1]. As a result, a common practice is to transfer graphs into vectors [5–9] in structure space or in Euclidean space, so that traditional machine learning algorithms such as Support Vector Machines (SVM) and Decision Tree can be applied. In the structure space (also referred to as quotient space) [7,8], the distance relations and nature of the original data are preserved, and some geometrical and analytical concepts such as derivatives of functions on structures can be determined, so that it can be applied to solve problems in structural pattern recognition. In the Euclidean space, the structural relations may be lost, but it provides simpler and more powerful analytical techniques for data analysis. Therefore, numerous approaches [10,9,11–18] have been proposed to represent graphs in Euclidean space. The key idea of transferring graphs into vectors in Euclidean space is to extract a set of subgraphs as features and use the presence/absence of features to represent each graph. From a feature selection perspective [19], these subgraph-based algorithms follow a filter approach for graph classification, i.e., the subgraph feature selection and the subsequent model training are separated into two steps. In summary, existing filter-based graph classification methods roughly fall into the following two categories.

* Correspondence to: Centre for Quantum Computation & Intelligent Systems, FEIT, University of Technology, Sydney, Ultimo, NSW 2007, Australia. Tel.: +61 450768511, fax: +61 2 9514 4535.
E-mail addresses: shirui.pan@uts.edu.au (S. Pan), jia.wu@student.uts.edu.au (J. Wu), xzhu3@fau.edu (X. Zhu), guodong.long@uts.edu.au (G. Long), chengqi.zhang@uts.edu.au (C. Zhang).

[1] In this paper, we only consider graphs with labels but no other feature values on nodes and edges.

http://dx.doi.org/10.1016/j.patcog.2015.05.019
0031-3203/

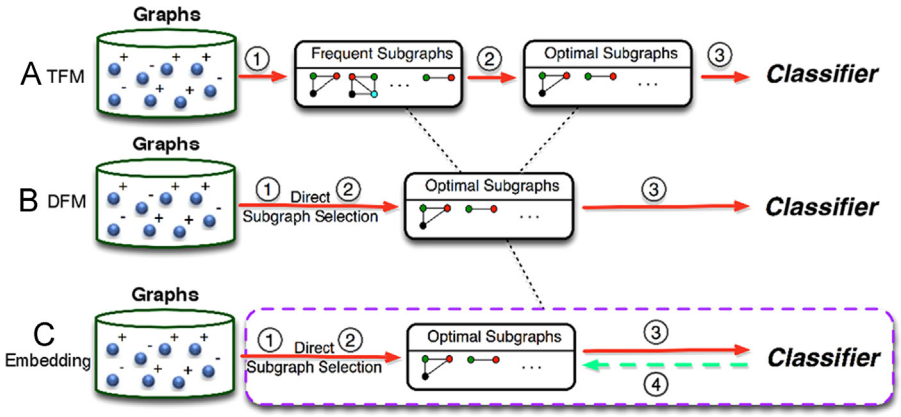

Fig. 1. Subgraph-based methods for graph classification from the feature selection perspective. TFM methods (A) sequentially perform frequent subgraph mining ①, optimal feature selection ②, and classifier learning process ③. DFM methods (B) integrate the feature selection ② into the frequent subgraph mining ① process. Our embedding method RLMD (C) unifies all steps (①②③) into a whole framework, and iterates until convergence ④.

Two-step Filter Methods (TFMs): This type of method first mines a set of frequent subgraphs as features and then applies a feature selection procedure to the discovered subgraphs, and uses the selected subgraph features to learn a classifier (*e.g.*, an SVM or Naive Bayes), as shown in Fig. 1(A). An early study [9] has shown that using frequent subgraphs as features can achieve reasonable good classification results. However, because TFMs separate subgraph feature discovery and feature evaluation into two steps, they may suffer from severe disadvantage in that the number of discovered subgraphs will grow exponentially when the minimum support value for subgraph mining is low. As a result, it will make the feature selection step heavily time-consuming. On the other hand, for relatively high minimum support values, many good subgraphs are pruned out because they do not meet the frequency requirement, so cannot be found to represent graphs.

Direct Filter Methods (DFMs): To improve the subgraph feature selection efficiency, numerous approaches [11,12,15–17] have been proposed to combine subgraph mining and feature selection into one step, representing a *direct discriminative feature selection* [18] scheme. So the feature selection is integrated into a subgraph mining process (Fig. 1 (B)), with pruning rules derived from the anti-monotone property of the significance (p-value) of each graph being used to reduce the search space. While DFMs substantially overcome the subgraph feature selection bottleneck, they also have a number of major disadvantages: (1) The subgraph selection is separated from the model learning process, so the selected subgraphs features may not best fit the underlying learning model, and (2) all these methods require users to specify the number of subgraph features K, whereas the optimal number of subgraphs K required for training a good classifier for graph classification is unknown and difficult to determine. Although subgraphs are selected using optimized measures, due to the redundancy inside the feature set, the accuracy of the classifiers, when varying the number of selected subgraph features K, is highly variable, as shown in Fig. 2. This is a common problem for all existing filter-based graph classification methods.

The above observations motivate the proposed research which *aims* to *integrate* subgraph mining, feature selection, and model training into one single framework (Fig. 1(C)) with the optimal number of subgraphs K being automatically determined for graph classification. To achieve this goal, we formulate subgraph feature selection as the problem of learning a K-dimensional feature space from a *huge subgraph space* in order to result in minimum regularized loss on the training data as follows:

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \gamma R(\boldsymbol{w}) \tag{1}$$

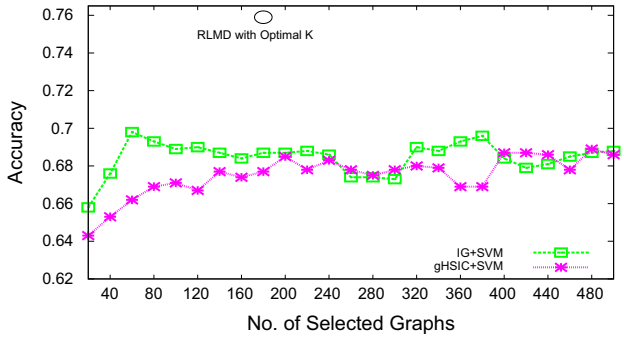

Fig. 2. Classification accuracy for filter subgraph-based methods *w.r.t.* different numbers of subgraphs on the NCI-1 chemical compound dataset. IG is a TFM method which uses information gain to select subgraphs, whereas gHSIC [12] is a DFM method. All methods use SVM as a base classifier. The optimal number of subgraph features K is crucial, but difficult to decide for filter methods. In comparison, the proposed method (RLMD) automatically finds 180 best subgraphs and achieves the highest accuracy, which is 6% more accurate than the second best method.

where $\{\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$ are the vector representations of the training graphs, $\mathcal{L}$ is a loss function measuring the difference between the prediction $f(\boldsymbol{x}_i)$ and the true label y_i, and $R(\boldsymbol{w})$ is a regularization term on parameters $\boldsymbol{w}$ to avoid over-fitting.

Indeed, the optimization in (1) has been widely studied [20–22] in machine learning community, but mainly for data with vector format. Several significant challenges remain for graph data:

1. *Implicit Subgraph Features:* For graph classification, no subgraph features are readily available (i.e., $\boldsymbol{x}_i$ is unknown) for training the model in (1). Instead, the feature space used to represent graphs is implicit and needs to be discovered by subgraph mining procedure as needed.
2. *K-dimensional Features from Huge Subgraph Space:* The number of subgraph candidates representing graphs is exponentially large. Finding an optimal number of K subgraphs for different graph datasets (in order to result in best classifiers), is crucial but has not been addressed by existing research.

In this paper, we propose a *unified* regularized loss minimization-driven (RLMD) graph classification framework. Our theme is to progressively select the most discriminative subgraph features from the training data in order to achieve minimum regularized loss for a well defined objective function. To integrate subgraph selection into the model learning process (*Challenge 1*), we formulate an objective function and design a subgradient method