Formulation of Learning Algorithms for Graph Data Classification

A thesis submitted

in partial fulfillment for the award of the degree of

Doctor of Philosophy

by

Asif Salim

Department of Mathematics

Indian Institute of Space Science and Technology

Thiruvananthapuram, India

January 2023

Abstract

The data in the forms of graphs are crucial in many domains of science and technology like bio-informatics, chemo-informatics, social media analysis, natural language processing, etc. These domains create graphs in a variety of forms. For example, a collection of graphs is created in bio-informatics and chemo-informatics domains where we need to do a conventional classification or regression. The nodes in these graphs may be represented as atoms in a molecule and edges represent the interaction between them. The nodes and edges can be accompanied with a discrete label or some attributes in the form of a vector. In social media and natural language processing, the graph is in the form of a single large network where we need to process on the nodes and edges. This thesis discusses the design of techniques that process different types of graph data by making use of tools in kernel methods and graph signal processing. Although the algorithms are discussed in the context of classification, with appropriate changes in the learning algorithms, they are also applicable for regression problems.

Our first approach is to make use of graph embedding techniques and multiple kernel learning (MKL). The graph embedding is the process of representing the graph in a vector space using properties of the graph while MKL is a framework where the optimal kernel is learned as a linear combination of a set of base kernels. The technique of MKL allows us to incorporate multiple graph embedding into a single learning framework. Hence we designed graph embedding using a multi-view approach, where each view is an embedding of the graph using a graph property. The reproducing kernel used in SVM is represented as a linear combination of the kernels defined on the individual embeddings. The proposed method helps to process a dataset of a collection of graphs with a categorical label information over the nodes and edges.

In the second approach, we made use of the optimal assignment kernel framework to design graph kernels. The bijection associated with the optimal assignment framework is defined between sets that consist of the nodes of the graph kernel arguments. In the proposed kernels, the nodes of the given data are divided into groups named as *neighbourhood sets* on the basis of the labels generated by the Weisfeiler-Lehman (WL) test for graph isomorphism and a matrix representation is defined for them. A kernel is then defined over the domain that consists of the *neighbourhood sets* in terms of the matrix and an aggregate

measure of those kernel values is used for defining the kernels. The proposed kernels can be used for analyzing a collection of graphs with categorical labels on the nodes/edges and it can also be extended to the case of attributed graphs in which apart from the labels, the nodes also contain vector information.

The third approach is specifically proposed for attributed graphs. We formulated the design of a reproducing kernel suitable for processing the attributes, in which the similarity between two graphs is defined on the basis of neighborhood information of the graph nodes with the aid of a product graph formulation. We represent the proposed kernel as the weighted sum of two other kernels of which one is an R-convolution kernel that processes the attribute information of the graph and the other is an optimal assignment kernel that processes label information. They are formulated in such a way that the edges processed as part of the kernel computation have the same neighborhood properties and hence the kernel proposed makes a well-defined correspondence between regions processed in the graphs. We found that the kernel value of the argument graphs in each iteration of the WL algorithm can be obtained recursively from the product graph formulated in our method.

The fourth approach is developed to classify the nodes of a single large network in contrast to the previous approaches. The spectral graph convolutional neural networks (SGCN) are utilized for this. In this work, it has been identified that the filters in the state-of-the-art SGCNs are essentially graph kernels in the form of low pass filters. They enforce a smoothness across the graph and use the functions of graph Laplacian as a tool that injects graph structure into the learning algorithm. The existing SGCNs are reviewed in the context of the relationship between graph Laplacian and regularization operators and propose a framework where the state-of-the-art filter designs can be deduced as its special cases. A new set of filters are designed that are associated with a well-defined low pass behavior. We also deduce the connection of support vector kernels and SGCN filters based on our framework.

The efficiency of the first three approaches are evaluated by incorporating the proposed kernels on support vector machines for classification task and the fourth approach on neural networks for semi-supervised node classification task on the real-world data sets and they have shown superior performance in comparison with that of the state-of-the-arts.