Chapter 11

GRAPH CLASSIFICATION

Koji Tsuda
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST)
Tokyo, Japan
koji.tsuda@aist.go.jp

Hiroto Saigo
Max Planck Institute for Informatics
Saarbrucken, Germany
hiroto@mpi-inf.mpg.de

Abstract Supervised learning on graphs is a central subject in graph data processing. In graph classification and regression, we assume that the target values of a certain number of graphs or a certain part of a graph are available as a training dataset, and our goal is to derive the target values of other graphs or the remaining part of the graph. In drug discovery applications, for example, a graph and its target value correspond to a chemical compound and its chemical activity. In this chapter, we review state-of-the-art methods of graph classification. In particular, we focus on two representative methods, graph kernels and graph boosting, and we present other methods in relation to the two methods. We describe the strengths and weaknesses of different graph classification methods and recent efforts to overcome the challenges.

Keywords: graph classification, graph mining, graph kernels, graph boosting

1. Introduction

Graphs are general and powerful data structures that can be used to represent diverse kinds of objects. Much of the real world data is represented not

C.C. Aggarwal and H. Wang (eds.), *Managing and Mining Graph Data*,
Advances in Database Systems 40, DOI 10.1007/978-1-4419-6045-0_11,

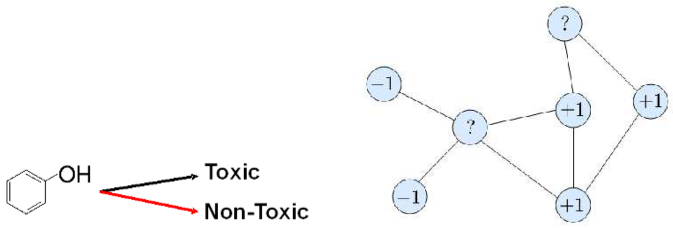

Figure 11.1. Graph classification and label propagation.

as vectors, but as graphs (including sequences and trees, which are specialized graphs). Examples include biological sequences, semi-structured texts such as HTML and XML, chemical compounds, RNA secondary structures, API call graphs, etc. The topic of graph data processing is not new. Over the last three decades, there have been continuous efforts in developing new methods for processing graph data. Recently we have seen a surge of interest in this topic, fueled partly by new technical advances, for example, development of graph kernels [21] and graph mining [52] techniques, and partly by demands from new applications, for example, chemical informatics. In fact, chemical informatics is one of the most prominent fields that deal with large repositories of graph data. For example, NCBI's PubChem has millions of chemical compounds that are naturally represented as molecular graphs. Also, many different kinds of chemical activity data are available, which provides a huge test-bed for graph classification methods.

This chapter aims at giving an overview of existing graph classification methods. The term "graph classification" can mean two different tasks. The first task is to build a model to predict the class label of a whole graph (Figure 11.1, left). The second task is to predict the class labels of nodes in a large graph (Figure 11.1, right). For clarity, we used the term to represent the first task, and we call the second task "label propagation"[6]. This chapter mainly deals with graph classification, but we will provide a short review of label propagation in Section 5.

Graph classification tasks can either be unsupervised or supervised. Unsupervised methods classify graphs into a certain number of categories by similarity [47, 46]. In supervised classification, a classification model is constructed by learning from training data. In the training data, each graph (e.g., a chemical compound) has a target value or a class label (e.g., biochemical activity). Supervised methods are more fundamental from a technical point of view, because unsupervised learning problems can be solved by supervised methods via probabilistic modeling of latent class labels [46]. In this chapter, we focus on two supervised methods for graph classification: graph kernels and graph boosting [40], which are similarity- and feature-based respectively. The two