Chapter 9
Mining Chemical Compounds

Mukund Deshpande, Michihiro Kuramochi, and George Karypis

Summary

In this chapter we study the problem of classifying chemical compound datasets. We present a substructure-based classification algorithm that decouples the substructure discovery process from the classification model construction and uses frequent subgraph discovery algorithms to find all topological and geometric substructures present in the dataset. The advantage of this approach is that during classification model construction, all relevant substructures are available allowing the classifier to intelligently select the most discriminating ones. The computational scalability is ensured by the use of highly efficient frequent subgraph discovery algorithms coupled with aggressive feature selection. Experimental evaluation on eight different classification problems shows that our approach is computationally scalable and on the average outperforms existing schemes by 10% to 35%.

9.1 Introduction

Discovering new drugs is an expensive and challenging process. Any new drug should not only produce the desired response to the disease but should do so with minimal side effects and be superior to the existing drugs on the market. One of the key steps in the drug design process is to identify the chemical compounds (widely referred to as *"hit"* compounds) that display the desired and reproducible behavior against the disease [247] in a biological experiment. The standard technique for discovering such compounds is to evaluate them with a biological experiment, known as an assay. The 1990s saw the widespread adoption of high-throughput screening (HTS), which

uses highly automated techniques to conduct the biological assays and can be used to screen a large number of compounds. Though in principle HTS techniques can be used to test each compound against every biological assay, it is never practically feasible for the following reasons. First, the number of chemical compounds that have been synthesized or can be synthesized using combinatorial chemistry techniques is extremely large. Evaluating this large set of compounds using HTS can be prohibitively expensive. Second, not all biological assays can be converted to high-throughput format. Third, in most cases it is hard to find all the desirable properties in a single compound, and chemists are interested in not just identifying the hits but studying what part of the chemical compound leads to desirable behavior so that new compounds can be rationally synthesized.

The goal of this chapter is to develop computational techniques based on classification that can be used to identify the hit compounds. These computational techniques can be used to replace or supplement the biological assay techniques. One of the key challenges in developing classification techniques for chemical compounds stems from the fact that the properties of the compounds are strongly related to their chemical structure. However, traditional machine learning techniques are suited to handling datasets represented by multidimensional vectors or sequences and cannot handle the structural nature of the chemical structures.

In recent years two classes of techniques have been developed for solving the chemical compound classification problem. The first class builds a classification model using a set of physicochemical properties derived from the compounds structure, called quantitative structure-activity relationships (QSAR) [14, 167, 168], whereas the second class operates directly on the structure of the chemical compound and in the attempt to automatically identify a small number of chemical substructures that can be used to discriminate between the different classes [41, 99, 193, 236, 436]. A number of comparative studies [222, 384] have shown that techniques based on the automatic discovery of chemical substructures are superior to those based on QSAR properties and require limited user intervention and domain knowledge. However, despite their success, a key limitation of these techniques is that they rely on heuristic search methods to discover these substructures. Even though such approaches reduce the inherently high computational complexity associated with these schemes, they may lead to suboptimal classifiers in cases in which the heuristic search failed to uncover substructures that are critical for the classification task.

In this chapter we present a substructure-based classifier that overcomes the limitations associated with existing algorithms. One of the key ideas of this approach is to decouple the substructure discovery process from the classification model construction step and use frequent subgraph discovery algorithms to find all chemical substructures that occur a sufficiently large number of times. Once the complete set of these substructures has been