On Association Rule Mining for the QSAR Problem

Luminita Dumitriu
"Dunarea de Jos" University, Romania

Cristina Segal
"Dunarea de Jos" University, Romania

Marian Craciun
"Dunarea de Jos" University, Romania

Adina Cocu
"Dunarea de Jos" University, Romania

INTRODUCTION

The concept of Quantitative Structure-Activity Relationship (QSAR), introduced by Hansch and co-workers in the 1960s, attempts to discover the relationship between the structure and the activity of chemical compounds (SAR), in order to allow the prediction of the activity of new compounds based on knowledge of their chemical structure alone. These predictions can be achieved by quantifying the SAR.

Initially, statistical methods have been applied to solve the QSAR problem. For example, pattern recognition techniques facilitate data dimension reduction and transformation techniques from multiple experiments to the underlying patterns of information. Partial least squares (PLS) is used for performing the same operations on the target properties. The predictive ability of this method can be tested using cross-validation on the test set of compounds.

Later, data mining techniques have been considered for this prediction problem. Among data mining techniques, the most popular ones are based on neural networks (Wang, Durst, Eberhart, Boyd, & Ben-Miled, 2004) or on neuro-fuzzy approaches (Neagu, Benfenati, Gini, Mazzatorta, & Roncaglioni, 2002) or on genetic programming (Langdon, &Barrett, 2004). All these approaches predict the activity of a chemical compound, without being able to explain the predicted value.

In order to increase the understanding on the prediction process, descriptive data mining techniques have started to be used related to the QSAR problem. These techniques are based on association rule mining.

In this chapter, we describe the use of association rule-based approaches related to the QSAR problem.

BACKGROUND

Association rule mining, introduced by (Agrawal, Imielinski & Swami, 1993), is defined as finding all the association rules between sets of items in a database that hold with more that a user-given minimum support threshold and a user-given minimum confidence threshold. According to (Agrawal, Imielinski & Swami, 1993) this problem is solved in two steps:

1. Finding all frequent itemsets in the database.
2. For each frequent itemset I, generating all association rules $I' \Rightarrow I \setminus I'$, where $I' \subset I$.

The second problem can be solved in a straightforward manner after the first step is completed. Hence, the problem of mining association rules is reduced to the problem of finding all frequent itemsets. This is not a trivial problem, since the number of possible frequent itemsets is equal to the size of the power set of I, $2^{|I|}$.

There are many algorithms proposed in the literature, most of them based on the Apriori mining method (Agrawal & Srikant, 1994) that relies on a basic property of frequent itemsets: all subsets of a frequent itemset are frequent. This property can also be stated as all supersets of an infrequent itemset are infrequent. There are other approaches, namely the closed-itemset approaches, as Close (Pasquier, Bastide, Taouil & Lakhal, 1999),

CHARM (Zaki & Hsiao, 1999) and Closet (Pei, Han & Mao, 2000). The closed-itemset approaches rely on the application of Formal Concept Analysis to association rule problem that was first mentioned in (Zaki & Ogihara, 1998). For more details on lattice theory see (Ganter & Wille, 1999). Another approach leading to a small number of results is finding representative association rules (Kryszkiewicz, 1998).

The difference between Apriori-based and closed itemset-based approaches consists in the treatment of sub-unitary confidence and unitary confidence association rules, namely Apriori makes no distinction between them, while FCA-based approaches report sub-unitary association rules (also named partial implication rules) structured in a concept lattice and, eventually, the pseudo-intents, a base on the unitary association rules (also named global implications, exhibiting a logical implication behavior). The advantage of a closed itemset approach is the smaller size of the resulting concept lattice versus the number of frequent itemsets, *i.e.* search space reduction.

MAIN THRUST OF THE CHAPTER

While there are many application domains for the association rule mining methods, they have only started to be used in relation to the QSAR problem. There are two main approaches: one that attempts classifying chemical compounds, using frequent sub-structure mining (Deshpande, Kuramochi, Wale, & Karypis, 2005), a modified version of association rule mining, and one that attempts predicting activity using an association rule-based model (Dumitriu, Segal, Craciun, Cocu, & Georgescu, 2006).

Mined Data

For the QSAR problem, the items are called chemical compound descriptors. There are various types of descriptors that can be used to represent the chemical structure of compounds: chemical element presence in a compound, chemical element mass, normalized chemical element mass, topological structure of the molecule, geometrical structure of the molecule etc. Generally, a feature selection algorithm is applied before mining, in order to reduce the search space, as well as the model dimension. We do not focus on feature selection methods in this chapter.

The classification approach uses both the topological representation that sees a chemical compound as an undirected graph, having atoms in the vertices and bonds in the edges and the geometric representation that sees a chemical compound as an undirected graph with 3D coordinates attached to the vertices.

The predictive association-based model approach is applied for organic compounds only and uses typical sub-structure presence/count descriptors (a typical substructure can be, for example, - CH_3 or –CH_2-). It also includes a pre-clustered target item, the activity to be predicted.

Resulting Data Model

The frequent sub-structure mining attempts to build, just like frequent itemsets, frequent connected sub-graphs, by adding vertices step-by step, in an Apriori fashion. The main difference from frequent itemset mining is that graph isomorphism has to be checked, in order to correctly compute itemset support in the database. The purpose of frequent sub-structure mining is the classification of chemical compounds, using a Support Vector Machine-based classification algorithm on the chemical compound structure expressed in terms of the resulted frequent sub-structures.

The predictive association rule-based model considers as mining result only the global implications with predictive capability, namely the ones comprising the target item in the rule's conclusion. The prediction is achieved by applying to a new compound all the rules in the model. Some rules may not apply (rule's premises are not satisfied by the compound's structure) and some rules may predict activity clusters. Each cluster can be predicted by a number of rules. After subjecting the compound to the predictive model, it can yield:

- a "none" result, meaning that the compound's activity can not be predicted with the model,
- a cluster id result, meaning the predicted activity cluster,
- several cluster ids; whenever this situation occurs it can be dealt with in various manners: a vote can be held and the majority cluster id can be declared a winner, or the rule set (the model) can be refined since it is too general.