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 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 than 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 \( I \Rightarrow I' \), where \( I' \subseteq I \).

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),