# Towards prediction of structured values

Pieter-Jan Drouillon,Hendrik Blockeel

Dept. of Computer Science, KULeuven, Belgium
{Pieter-Jan.Drouillon,Hendrik.Blockeel}@cs.kuleuven.be

In the field of data mining, learning predictive models is a common task. Based on input values, a predictive model delivers a set of output values. These input values are a mixture of different types such as numerical, categorical or structured values. They are entered in the model which computes the outcome. The output is typically a single value or a vector of values.

Up till now, the most common output values are numerical or categorical values. In the past, some attempts were made to predict structured values[1]. The focus of our research is to investigate what the needs and possible restrictions are to include structured values as output.

Among many applications of this type of prediction, the task of deriving the structure of a molecule solely based on its mass spectrogram is an example.This procedure is often done to identify an unknown compound. In mass spectroscopy, molecules of a compound are bombarded with electrons. Some break up to give a variety of charged fragments, characteristic of the original molecule. A mass spectrogram is basically a graph of the mass to charge ratio of the different fragments versus the frequency. So the input values are numerical values, the output is the structure of the original molecule.

For this example application, a database of mass spectrograms has been collected[2]. It contains the name, the chemical formula, the mass and the mass spectrogram of each molecule. This database will first be used to predict the number of occurrences of each different element type in the molecule. This is a strictly simpler task which can be used as a reference point to evaluate the developed algorithms against. In a second phase, we will investigate the possible use of learning methods that handle structured values, such as Inductive Logic Programming. A possible approach is to predict numerical and categorical features of the molecule such as types of bonds or number of functional groups. Combined with constraints, functional groups can be suggested.

## References

[1] Jan Ramon and Luc De Raedt. Instance Based Function Learning *Lecture Notes in Computer Science*,volume 1634,1999

[2] SDBS, National Institute of Advanced Industrial Science and Technology