

On June 3, 2002 a **'Symposium from Machine Learning to Data Mining'** was organised by the Machine Learning group at Leuven University. 63 participants from Belgium and the Netherlands witnessed 4 invited talks and finally Wim Van Laer's PhD defence. A summary follows.

The first presentation of the day entitled **'Inductive databases: a declarative data mining approach'** was given by Luc De Raedt. The philosophy of the presented approach was to make 'first-class citizen' of patterns by thinking of them as data, stored in an inductive database. Data mining now becomes querying this database for patterns that satisfy certain constraints. A first framework was given for monotonic constraints, such as those based on frequency or generality. It is clear that more work needs to be done on non-monotonic constraints, such as accuracy, etc.

An interesting talk, entitled **'Building and mining the multidimensional HIV data cube'** was given by Elke Van Craenenbroeck and Luc Dehaspe from PharmaDM, Leuven, Belgium. The talk addressed the issue of mining from multidimensional data, in a sense to automate some of the interactive data analysis from OLAP. The proposed method was to use a star or snowflake schema to represent the datacube in relational form, and then apply a multi-relational learner. Such schemata are interesting because they are on the border of propositional and multi-relational data mining: yes, there are multiple tables, but all relations with the central fact table are determinate, so a single join can be produced without loss of information (denormalisation). The speakers opted for a multi-relational approach because the more compact data representation would be more efficient.

The method was applied to a test database (2 dimensions) of HIV data. Each cell in the datacube basically represented the outcome (0 or 1) of a particular experiment on a particular strain. By 'rolling up' over the available hierarchies in the two dimensions, concept hierarchies could be generated.

A large part of Peter Flach's presentation **'Descriptive data mining: current issues'** was dedicated to the introduction of basic concepts and techniques for descriptive data mining. These include individual centred representations (for structured data), subgroup discovery, rule evaluation, confusion matrices and ROC analysis applied to rules. As such, the presentation can be recommended as a good introduction to important concepts in descriptive data mining.

A number of new results were presented. First there are two new rule evaluation measures, Satisfaction and Confirmation, for which good bounds can be given to prune the search space. Second an improvement of CN2, called CN2-SD was given which assigned weights to examples, rather than remove them after each rule is applied.

The last invited speaker, Saso Dzeroski, gave a talk entitled **'Is combining Classifiers Better than Selecting the Best One?'**. The talk addressed the issue of how to best combine classifiers (ensemble) and specifically of how to learn to best combine (stacking). A thorough experimental survey seems to indicate that most existing methods do not outperform the simple approach of selecting the best classifier by cross-validation, in a sense answering the question in the title negatively. The speaker however presented one method, multi-response model trees, that does have a clear edge on selecting the best classifier.

The final event of the day, and in fact the motivation for organising the seminar was the PhD defence of (now Dr.) Wim Van Laer. He did a very good job of explaining his work in understandable terms to the many friends and relatives who showed up, although some of the finer points of theta-subsumption may have been lost to one or two uncles. The thesis was well received by the jury, and the much-desired title obtained.

Report written by Arno Knobbe, Kiminkii, the Netherlands.