Use of statistical testing for determining the values of parameters for an unsupervised feature construction algorithm

Marian-Andrei Rizoiu*
ERIC Laboratory, University Lyon2, France Marian-Andrei.Rizoiu@univ-lyon2.fr

Julien Velcin
ERIC Laboratory, University Lyon2, France Julien.Velcin@univ-lyon2.fr

Stéphane Lallich
ERIC Laboratory, University Lyon2, France Stephane.Lallich@univ-lyon2.fr

Data Mining algorithms are powerful tools for extracting knowledge from raw data, but often suffer from having a plethora of parameters. Usually, these parameters are dependent on the dataset and require the human user to tune them manually. In this presentation, we will propose a method to address this problem. This method uses statistical testing and replaces the data-dependent parameter with a data-independent significance level. We will apply this method to a feature construction algorithm, which tries to catch the underlying semantic structure of a feature set by replacing highly correlated pairs of features with conjunction of the primitive features or their negations. We show that this parameter selection method obtains results comparable to those of a brute force method in which parameters are varied and the best solution is selected afterwards.

Key Words: statistical tests, data mining, parameter determination, feature construction.