

Data Mining Scenarios for the Discovery of Subtypes and the Comparison of Algorithms

Fabrice Colas
Leiden University, Netherlands
<http://www.liacs.nl/~fcolas/>

Abstract:

A data mining scenario is a logical sequence of steps to infer patterns from data. In this talk, we will present two of them with a particular emphasis on the second one.

The first scenario aims to identify homogeneous subtypes in cohorts of patients affected by diseases presenting clinical heterogeneity (e.g. Osteoarthritis, Parkinson's disease). It enabled to classify the patients more sensitively and therefore, contributed to the search for the underlying mechanisms of the diseases. When applied to databases of drug compounds, this scenario aimed to improve our understanding of the similarity (and distance) between the different phenotypic effects induced by drugs and chemicals.

Our second scenario aims to compare text classification algorithms. First, we show that common classifiers achieve comparable performance on most problems. Second, tightly constrained SVM solutions are high performers. In that situation, most training documents are bounded support vectors, SVM reduces to a nearest mean classifier and no training is necessary, which raises a question on SVM merits in sparse bag of words feature spaces. Also, SVM is shown to suffer from performance deterioration for particular combinations of training set size/number of features. This relate to outlying documents of distinct classes overlapping in the feature space.