Mining Sequential Patterns

by Jilles Vreeken

http://people.mmci.uni-saarland.de/~jilles/

Pattern mining is a powerful tool for exploratory data analysis, aimed at identifying interesting *local* structure in data -- as opposed to fitting a global model, such as topic models. What a pattern is depends on the pattern language. For text, for example, subsequences are natural patterns, -- for which we can consider many variations, univariate or multivariate, with or without gaps, with or without local choices, etc, etc.

The goal of pattern mining is to discover *interesting* patterns. Unsurprisingly, this is hard. The traditional approach, for example, is not very useful: by answering the question 'find me all potentially interesting patterns' typically far too many results are returned -- and many of which will be redundant. Instead, we take an information theoretic approach, and ask for the set of patterns that describes your data best. This set has many desirable properties: it is small, captures the most important structure in your data, while being neither redundant nor overfit.

I will give examples of this approach for univariate and multivariate data. Example results will resp. include characteristic local word orders, as well as `translations' when we mine (aligned) text over different languages. Moreover, in a more general setting, I will show that these `patterns that compress' are indeed useful in a wide range of data mining tasks, including classification, anomaly detection, missing value estimation, and clustering, in which these pattern sets have been shown to obtain top-notch and highly interpretable results, without the need for parameters.