SEQUENCE CLASSIFICATION THROUGH RELEVANT SEQUENCE PATTERN MINING

Pedro Gabriel Dias Ferreira
pedrogabriel@di.uminho.pt
Department of Informatics,
University of Minho, Campus de Gualtar,
4710-057 Braga, Portugal

We tackle the problem of sequence classification using relevant subsequences found in a class labeled sequence dataset. A subsequence is *relevant* if it is frequent, maximal and with length greater or equal than 2. Classification is performed by assigning different scores to the relevant subsequences found in the unseen sequence. The score is calculated based on the weight attributed to the relevant subsequences and the percentage of classes where they occur. We propose two different score schemes: one is length based and the other is calculated by identifying the gaps occurring in the subsequence. The combination of these features results in a multi-class and multi-domain method that is exempt of data preprocessing and background knowledge. We illustrate the performance of our method using datasets of proteins, linux commands and synthetic generated data. The tests generally show promising results, even when the training data is small.