Kernel Methods and Algorithms for General Sequence Analysis

Pavel P. Kuksa, Pai-Hsi Huang, Vladimir Pavlovic
Rutgers University
Piscataway, NJ 08854
pkuksa@cs.rutgers.edu

May 8, 2008

Abstract

Problems of analysis and modeling of sequential data arise in many practical applications. In this work, we develop efficient algorithms and methods for general sequence analysis. In particular, we propose novel ways of modeling sequences under complex transformations (such as multiple insertions, deletions, mutations) and present a new family of similarity measures (kernels), spatial string kernels, that can be computed very efficiently and show state-of-the-art performance on a variety of distinct classification tasks. We also present new algorithms for approximate (e.g. with mismatches) string comparison that improve currently known time bounds for such tasks and show order-of-magnitude running time improvements. In an extensive set of experiments on many challenging classification problems, such as detecting homology (evolutionary similarity) of remotely related proteins, categorizing texts, and performing classification of music samples, proposed algorithms and measures display state-of-the-art classification performance and run substantially faster than existing methods. We solve these problems in both binary and multi-class settings, as well as apply our methods to large-scale datasets with partially labeled samples.