Rutgers Computer Science Technical Report RU-DCS-TR684
February 2011

Generalized Similarity Kernels for Efficient Sequence Classification

by

Pavel P. Kuksa, Imdadullah Khan, Vladimir Pavlovic
Rutgers University
Piscataway, NJ 08854
pkuksa@cs.rutgers.edu

ABSTRACT

String kernel-based machine learning methods have yielded great success in practical tasks of structured/sequential data analysis. In this paper we propose a novel computational framework that uses general similarity metrics and distance-preserving embeddings with string kernels to improve sequence classification. An embedding step, a distance-preserving bitstring mapping, is used to effectively capture similarity between otherwise symbolically different sequence elements. We show that it is possible to retain computational efficiency of string kernels while using this more "precise" measure of similarity. We then demonstrate that on a number of sequence classification tasks such as music, and biological sequence classification, the new method can substantially improve upon state-of-the-art string kernel baselines.