Pattern Discovery and Compression in Finite State Transducers

Giancarlo Schrementi
Indiana University

Michael Gasser
Indiana University

Abstract: Human languages are able to efficiently encode our environment because of their ability to recognize patterns in semantics and translate those patterns into a compressed syntax. This work describes evolving finite state transducers to efficiently encode their environment. The transducers are evolved using a genetic algorithm that makes use of the minimum description length (MDL) principal [1] to determine fitness. The MDL principle rewards transducers that produce compact, distinct encodings that the inverted form of the transducer is able to unambiguously decode. The environment that the transducers are encoding is a regular language that has structure for the transducers to discover and capture in their encodings. We show that this emphasis on successful decoding provides the genetic algorithm a strong incentive to find encodings that capture the structure inherent in the environment. The transducers are able to identify common components in the environment and distinctly represent them.

[1] Jorma Rissanen. Modeling by the shortest data description. Automatica, 14:465471, 1978.