Evolving Regular Expression Features for Text Classification with Genetic Programming  
R.J.P. Bakker

Abstract

Text classification algorithms often rely on vocabulary counters like bag-of-words or character n-grams to represent text as a vector appropriate for use in machine learning algorithms. In this work, automatically generated regular expressions are proposed as an alternative feature set. The proposed algorithm uses genetic programming to evolve a set of regular expression features based on labeled text data and train a classifier in an end-to-end fashion. Though a comparison of the generated features and traditional text features indicates a classifier using generated features is not able to make better predictions, the generated features are able to capture patterns that cannot be found with the traditional features. As a result, a classifier combining traditional methods with generated features is able to improve significantly.