MONTE CARLO TREE SEARCH FOR LARGE SCALE DATASET FEATURE SELECTION

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2012

By
Cristian Darío Quiroz Castellar
School of Computer Science

Abstract

MONTE CARLO TREE SEARCH FOR LARGE SCALE DATASET FEATURE SELECTION
Cristian Darío Quiroz Castellar
A dissertation submitted to the University of Manchester
for the degree of Master of Science, 2012

The problem of Feature Selection is a broadly studied one, consisting of selecting the measurements (or features) of a dataset that provide most information in order to reduce the complexity of a problem and it often carries great increase in performance when applying Machine Learning classification algorithms to the resultant *reduced* dataset. The importance of Feature Selection *cannot be stressed enough* [38], particularly when studying problems with large scaled data, where classic filter and wrapper Feature Selection approaches fall short to the task, being either too naive and simplistic or too computationally expensive. Monte Carlo Tree Search has been proposed as a Feature Selection algorithm, being an interesting candidate due to being designed to solve combinatorial problems in general, and the high amount of work done in research on it [9]. This dissertation builds and studies an implementation based on that algorithm, exploring optimizations as ways to reduce execution time, through the analysis of functions that help determining whether a subset of features will allow a reasonable accuracy after being used to train a classification model, called Reward Functions. Results show how a Reward Function based on the K-Nearest Neighbour classifier executes much faster than other proposed methods, maintaining similar performance in classification accuracy. This dissertation takes the first few steps to allow Monte Carlo Tree Search for Feature Selection work with large scale datasets and shows how it is a plausible goal with a wide variety of possibilities for optimization.