

Projeto de Algoritmos Baseados em Florestas de Posets para o Problema de Otimização U-curve

Instituto de Matemática e Estatística

Gustavo Estrela de Matos

September 18, 2017

Resumo

O problema U-curve é uma formulação de um problema de otimização que pode ser utilizado na etapa de seleção de características em Aprendizado de Máquina, com aplicações em desenho de modelos computacionais de sistemas biológicos. Não obstante, soluções propostas até o presente momento para atacar esse problema têm limitações do ponto de vista de consumo de tempo computacional e/ou de memória, o que implica na necessidade do desenvolvimento de novos algoritmos. Nesse sentido, em 2012 foi proposto o algoritmo Poset-Forest-Search (PFS), que organiza o espaço de busca em florestas de posets. Esse algoritmo foi implementado e testado, com resultados promissores; todavia, novos melhoramentos são necessários para que o PFS se torne uma alternativa competitiva para resolver o problema U-curve. Neste projeto propomos a construção de uma versão paralelizada e escalável do algoritmo PFS, utilizando diagramas de decisão binária reduzidos e ordenados. Além disso, propomos adaptar o PFS como um algoritmo de aproximação, no qual o critério de aproximação da solução ótima faça uso do teorema da navalha de Ockham. Os algoritmos desenvolvidos serão implementados e testados em instâncias artificiais e também em conjuntos de dados próprios para experimentos comparativos entre diferentes algoritmos de seleção de características.

Contents

1	Introduction	4
2	Conclusion	5

Chapter 1

Introduction

A seleção de características pode ser utilizada como um auxílio na construção de um modelo de aprendizado de máquina. Essa técnica consiste em, dado o conjunto de características observadas nas amostras, escolher um subconjunto que seja ótimo de acordo com alguma métrica. Devemos considerar a etapa de seleção de características na construção de um modelo de aprendizado quando a quantidade de características é muito grande, o que pode fazer o modelo ser muito caro computacionalmente; ou quando a quantidade de amostras é pequena comparada a complexidade do modelo original, em outras palavras, quando ocorre sobreajuste (do inglês, *overfitting*).

Chapter 2

Conclusion