Projetos de algoritmos baseados em florestas de posets para o problema de otimização U-Curve

Gustavo Estrela
com supervisão de Marcelo S. Reis

Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil
Center of Toxins, Immune-response and Cell Signaling (CeTICS), Instituto Butantan, Brazil
Laboratório Especial de Ciclo Celular (LECC), Instituto Butantan, Brazil

Motivação

Seleção de características é um problema de otimização combinatória em que, dado uma função custo c e um conjunto S, procura-se o $X \in \mathcal{P}(S)$ de custo mínimo. No contexto de aprendizado de máquina, a solução deste problema pode ser aplicada como uma ferramenta que diminui a complexidade de modelos, selecionando um subconjunto ótimo de atributos do objeto de classificação. O espaço de busca do problema induz o reticulado Booleano $(\mathcal{P}(S), \subseteq)$ e é comum que a função de custo descreva curvas em U nas cadeias desse reticulado, o que é explicado por erros de estimação que aumentam quando se adiciona características. Então, podemos aproximar este problema pelo problema U-curve. Existem algoritmos na literatura que exploram esta aproximação, entretanto as soluções atuais têm limitações de escalabilidade. Propomos neste trabalho estudar algoritmos atuais e criar novos algoritmos para solucionar o problema U-curve de forma mais eficiente.

Algoritmos de florestas

O algoritmo da literatura Poset-Forest-Search (PFS) representa o espaço de busca com duas florestas (Fig. 1). A busca pelo ótimo se dá pelo percorrimento de cadeias nestas florestas. A hipótese U-curve permite que hajam podas durante este percorrimento.

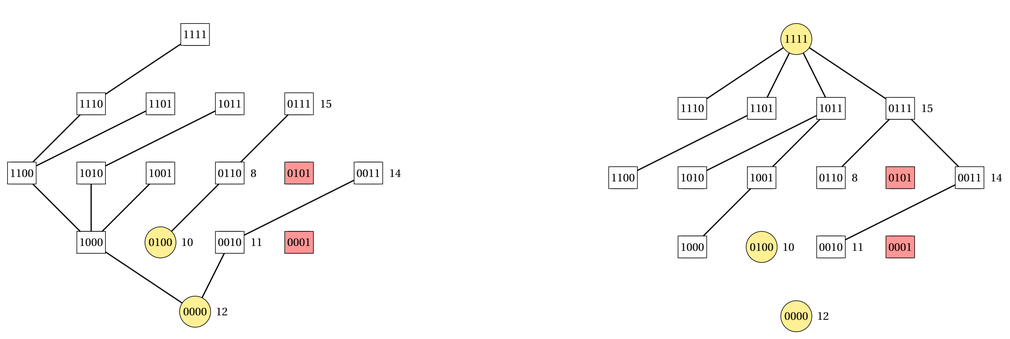


Fig. 1: Exemplo de florestas do algoritmo PFS. Os nós em vermelho foram removidos do espaço de busca, enquanto os nós em amarelo são raízes da floresta.

Implementamos uma variação do PFS que modifica o armazenamento e a escolha de raízes para percorrimentos na floresta, o ROBDD PFS (RPFS). Também para paralelizamos este algoritmo, criando o Parallel PFS (PPFS). Além disso, criamos um novo algoritmo paralelo que usa o U-Curve-Branch-and-Bound (UBB) para decompor a floresta em árvores menores que são resolvidas por chamadas independentes do PFS; este chamamos de UBB-PFS.

Parallel-U-Curve-Search

O Parallel-U-Curve-Search (PUCS) particiona o espaço de busca em estruturas que também são reticulados Booleanos. Esta divisão do espaço de busca permite aplicações recursivas do algoritmo e a solução paralela de cada parte. O algoritmo que resolve cada parte é chamado de algoritmo base.

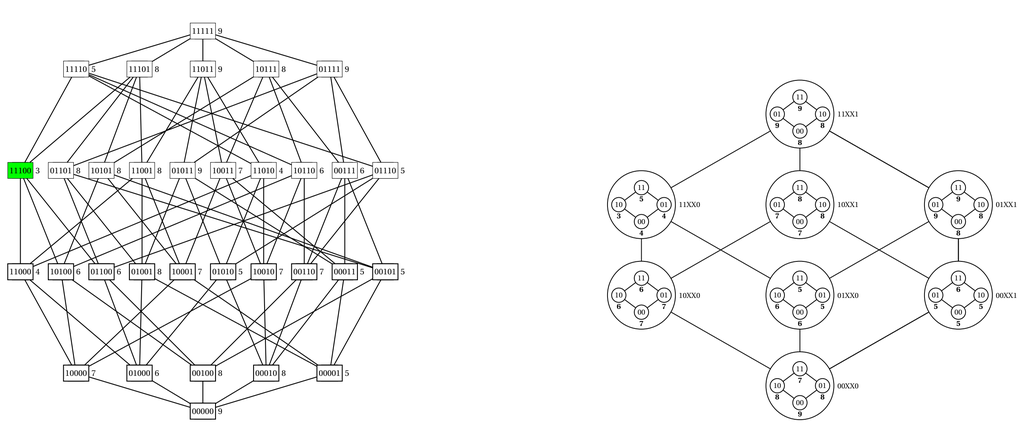


Fig. 2: Instância do problema U-curve com $|S| = 5$ (esquerda) e o particionamento feito quando a terceira de quarta variável são escolhidas como don't cares (direita).

Apoio Financeiro

FAPESP CNPq CAPES

Resultados

Usamos o *featsel* (github.com/msreis/featsel), um arcabouço em C++, para implementar e avaliar os novos algoritmos, comparando com soluções da literatura. Os testes foram conduzidos em uma servidora de 64 cores e 256 GB de memória RAM.

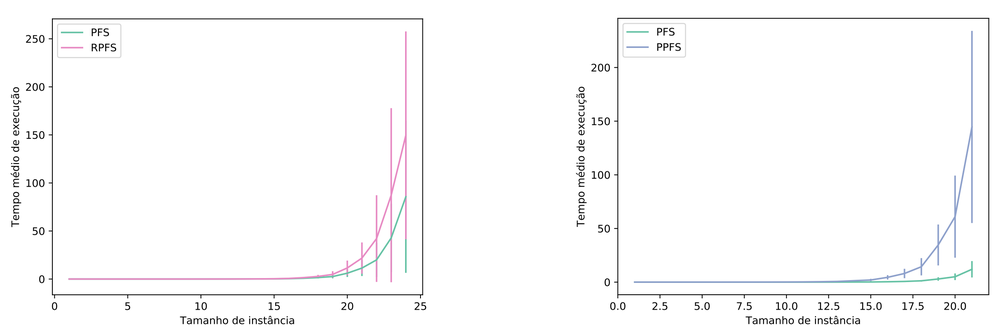


Fig. 3: Comparação do tempo médio de execução do PFS com as variações RPFS e PPFS.

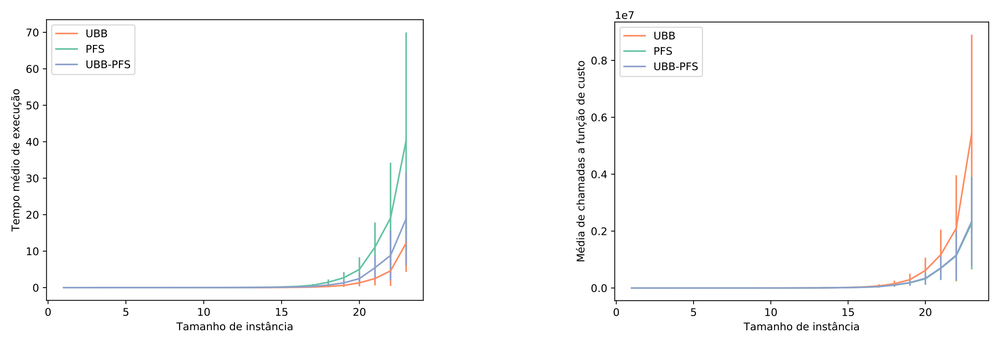
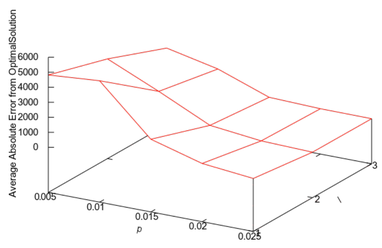


Fig. 4: Comparação do tempo médio de execução (esquerda) e número médio de chamadas da função custo (direita) feitas pelo UBB-PFS, comparando com o UBB e PFS.

O algoritmo PUCS possui dois parâmetros, p e l que definem a granularidade do particionamento. Quando o algoritmo base é uma heurística, notamos que estes parâmetros também controlam a qualidade da solução obtida pelo PUCS. Comparamos o PUCS heurístico com os algoritmos Sequential Forward Floating Selection (SFFS) e Best-First Search (BFS).



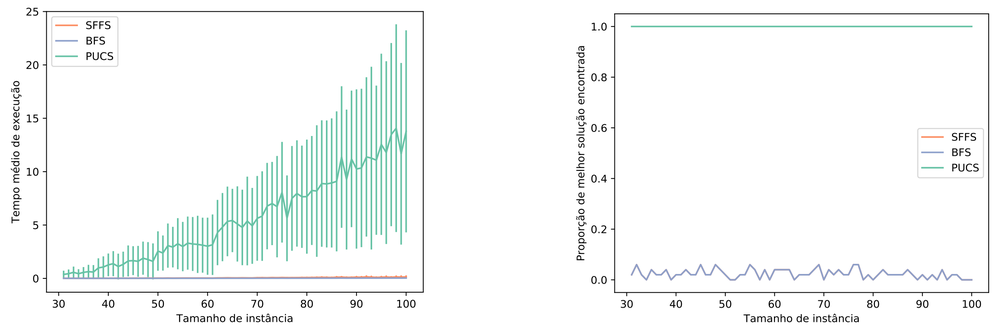


Fig. 5: Tempo de execução e proporção de melhor resposta encontrada pelos algoritmos sub-ótimos PUCS, SFFS e BFS.

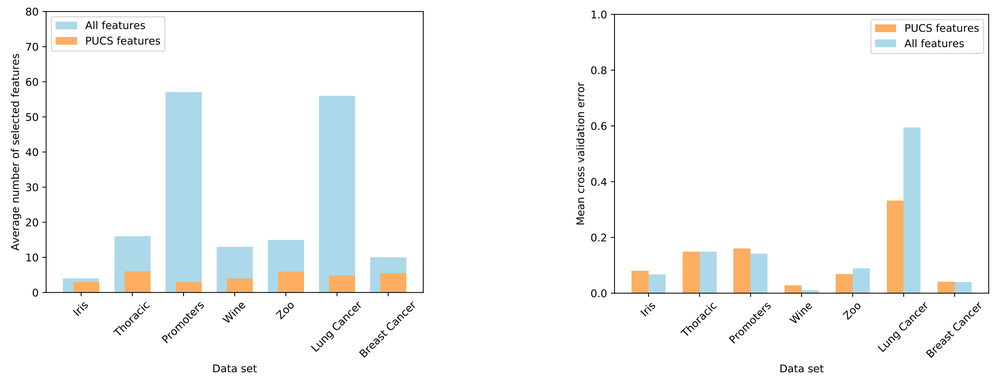


Fig. 6: Número médio de características selecionadas pelo PUCS e erro médio de validação cruzada quando utiliza-se as características selecionadas no projeto de classificadores do tipo SVM. Estes conjuntos de dados foram extraídos do UCI Machine Learning Repository.

Conclusão

Neste trabalho fomos capazes de criar algoritmos competitivos para o problema U-Curve e também confirmamos a eficácia da seleção de característiacs na seleção de modelos de aprendizado. Trabalhos futuros nesta linha incluem:

- Avaliação de robustez dos novos algoritmos quando a hipótese U-curve tem violações.
- Aplicação de seleção de características na identificação de vias de sinalização celular

{marcelo.reis, gustavo.matos }@butantan.gov.br