Relatório Científico Final – Iniciação Científica

Processo FAPESP 2016/25959-7

Projeto de algoritmos baseados em florestas de posets para o problema de otimização U-curve

Beneficiário: Gustavo Estrela de Matos

Responsável: Marcelo da Silva Reis

Relatório referente aos trabalhos desenvolvidos entre 1 de maio e 31 de dezembro de 2017

Laboratório Especial de Toxinologia Aplicada, Instituto Butantan

São Paulo, 4 de Janeiro de 2018

Conteúdo

1 Resumo do Projeto Proposto

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 modificações ao `PFS` na escolha de caminhos de percorrimento da floresta de busca, e na estrutura de dados utilizada para armazenar este grafo, com o uso de diagramas de decisão binária reduzidos e ordenados (OBDDs); também propomos a criação de uma versão paralela e escalável do algoritmo `PFS`. Além disso, propomos a criação de um algoritmo baseado no `PFS` que tenha características de um algoritmo de aproximação, no qual o critério de aproximação da solução ótima se baseie no teorema da navalha de Ockham. Os algoritmos desenvolvidos serão implementados no arcabouço *featsel* e testados com instâncias artificiais e também reais, com conjuntos de dados de aprendizado de máquina retirados do University of California Irvine (UCI) Machine Learning Repository.

2 Atividades Realizadas

2.1 Estudo de algoritmos baseados em florestas

O algoritmo `Poset-Forest-Search` (`PFS`) é um algoritmo ótimo para resolver o problema de otimização U-Curve e serviu de base para a criação da maioria dos algoritmos elaborados neste trabalho. O `PFS` é uma generalização de um outro algoritmo mais simples,

o `U-curve-Branch-and-Bound` (`UBB`), que é um algoritmo *branch-and-bound* ótimo que decompõe o espaço de busca em uma árvore, e acha o mínimo global do problema fazendo ramificações e podas nesta árvore.

A árvore de busca do `UBB` permite que a procura pelo mínimo ocorra de maneira parecida com uma busca em profundidade, que percorre cadeias do reticulado Booleano de subconjuntos menores para maiores. Sempre que o custo de um subconjunto X_i aumenta em comparação ao anterior X_j no percorrimento, a hipótese de que a função de custo é decomponível em curvas em U garante que a subárvore que começa em X_i pode ser removida do espaço de busca; chamamos este procedimento de poda. O algoritmo `UBB` tem, entretanto, uma limitação, pois quando a função de custo do problema é monótona não-crescente, a condição de poda nunca é verdadeira e o espaço de busca inteiro é visitado, o que compromete a escalabilidade do algoritmo.

O `PFS` enfrenta esta limitação ao fazer percorrimentos de cadeias do espaço de busca em duas direções, de conjuntos menores para maiores (como faz o `UBB`) e também o contrário. Para fazer isso, este algoritmo decompõe o espaço de busca em duas árvores, uma para cada direção de percorrimento. Com a criação de duas estruturas para representar o mesmo espaço de busca, torna-se necessário a atualização de uma estrutura sempre que a outra sofrer mudanças, e isto implica na utilização de florestas ao invés de árvores para representar o espaço de busca no `PFS`. Resumidamente, uma iteração do deste algoritmo deve escolher uma direção de percorrimento; fazer o percorrimento com poda (de maneira similar ao `UBB`); e, finalmente, atualizar a floresta dual a que foi percorrida.

Referências

[1] Reis, Marcelo S. "Minimization of decomposable in U-shaped curves functions defined on poset chains–algorithms and applications."PhD thesis, Institute of Mathematics and Statistics, University of São Paulo, Brazil, (2012).

[2] Reis, Marcelo S., Carlos E. Ferreira, and Junior Barrera. "The U-curve optimization problem: improvements on the original algorithm and time complexity analysis."arXiv preprint arXiv:1407.6067, (2014).

[3] Bryant, Randal E. "Graph-based algorithms for boolean function manipulation."IEEE Transactions on Computers, 100.8 (1986): 677-691.