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

Aluno: Gustavo Estrela de Matos

Orientador: Marcelo da Silva Reis

14 de Dezembro de 2016

Resumo

Design of poset forest-based algorithms for the U-curve optimization problem

Student: Gustavo Estrela de Matos

Supervisor: Marcelo da Silva Reis

14 de Dezembro de 2016

Abstract

Conteúdo

1 Introdução

1.1 O problema U-Curve

O problema de seleção de característica consiste em, dado um cojunto S de características, escolher um subconjunto de características que seja ótimo. A solução desse problema tem aplicações na construção de modelos para aprendizado de máquina e reconhecimento de padrões, que dependem da escolha de um subconjunto de características que seja o mais relevante possível (ótimo), de acordo com alguma métrica. Formalmente, podemos definir o problema de seleção de características como um problema de busca, no qual procuramos um subconjunto $X \in \mathcal{P}(S)$ que minimiza uma função de custo $c : \mathcal{P}(S) \to \mathbb{R}_+$.

O espaço de busca do problema de seleção de características pode ser visto como um reticulado booleano $(\mathcal{P}(S), \subseteq)$, onde cada nó é um conjunto de características, também chamado de classificador. É comum nesse problema que as cadeias do reticulado descrevam "curvas em u"quando avaliadas pela função de custo c. Esse comportamento pode ser intuitivamente explicado se considerarmos que um classificador melhora ao adicionarmos novas características até um ponto em que o grande número de características causa grandes erros de estimação, piorando o classificador.

O problema U-Curve é um caso particular do problema de seleção de características em que todas as cadeias do espaço de busca descrevem "curvas em u"quando avaliadas pela função de custo. Existem algoritmos ótimos para solução do problema U-Curve, como o Poset-Forest Search (`PFS`) e U-Curve Search (`UCS`) [1]. Além disso, em outra oportunidade de iniciação científica, estudamos o uso de diagramas de decisão binários ordenados e reduzidos (ROBDDs) como uma estrutura de dados eficiente para o controle do espaço de busca [2].

O uso de ROBDDs aliado a mudanças à dinâmica do `UCS` nos levaram a criação do algoritmo `UCSR`. Esse novo algoritmo trouxe melhoras no tempo de execução, pois permite consultas rápidas ao espaço de busca, o que era mais custoso no algoritmo `UCS`. Porém, as melhorias obtidas foram limitadas, uma vez que manter a estrutura de ROBDD, em alguns

casos, demandava grande processamento e uso de memória. Portanto, torna-se necessário a criação de novos algoritmos para resolver o problema, o que nos leva ao estudo do algoritmo Poset-Fores Search (`PFS`).

1.2 O algoritmo Poset-Forest Search

2 Objetivos

1. Utilização dos ROBDDs, implementados no IC anterior, para representar as listas de raízes do algoritmo PFS.

2. Desenho de uma versão paralelizada do PFS, com maior escalabilidade. Para este fim, paralelizaremos o percorrimento das florestas de posets, com o programa principal gerenciando a escolha das raízes (i.e., início de um percor- rimento), guardando o mínimo corrente e centralizando a atualização das podas.

3. Desenvolvimento de uma versão do PFS que funcione como algoritmo de aproximação para o problema U-curve, utilizando como critério de aproximação da solução ótima o teorema da navalha de Ockham:

   Dado um espaço de hipóteses H (i.e., espaço de busca), o número mínimo de amostras necessário para se obter uma solução que erra no máximo ϵ com $1-\delta$ de probabilidade é expresso por:

   $$m(\delta, \epsilon) = \frac{1}{\epsilon} log(\frac{|H|}{\delta}). \quad (1)$$

4. Implementação e testes dos algoritmos propostos, para isso empregando o arcabouço featsel.

3 Plano de trabalho

3.1 Cronograma

Tabela listando atividades de janeiro a dezembro de 2017.

3.2 Descrição de atividades

Descrição das atividades da tabela da subseção anterior.

4 Materiais e métodos

5 Forma de análise e divulgação de resultados

- Arcabouço featsel, já contando com os acréscimos de classes de ROBDDs;
- Biblioteca OpenMP.

6 Forma de Análise e de Divulgação dos Resultados

- Benchmarking contra outros algoritmos de seleção de caracteríscas;
- Elaboração de paper para ser enviado para publicação ao final da IC proposta.

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]