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

Instituto de Matemática e Estatística

Gustavo Estrela de Matos

11 de Outubro de 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.

Conteúdo

Capítulo 1

Introdução

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*).

Mais formalmente, o problema de seleção de características consiste em um problema de otimização combinatória em que, dado um conjunto S de características, procuramos por um subconjunto $X \in \mathcal{P}(S)$ ótimo de acordo com uma função de custo $c : \mathcal{P}(S) \to \mathbb{R}_+$. É comum nas abordagens do problema explorar o fato de que o espaço de busca $\mathcal{P}(S)$ junto a relação $\subseteq$ define um reticulado booleano. No geral, a função de custo c deve ser capaz de medir quão informativas as características X são em respeito ao rótulo Y do problema de aprendizado, portanto essa função costuma depender da estimação da distribuição de probabilidade de (X, Y).

Quando ocorre a estimação da distribuição de probabilidade conjunta de (X, Y), o custo das cadeias do reticulado booleano reproduzem um fenômeno conhecido em aprendizado de máquina, "curvas em u". Para entender intuitivamente esse fenômeno, devemos observar que conforme subimos uma cadeia do reticulado estamos aumentando o número de características sendo consideradas, portanto existem mais possíveis valores de X, permitindo descrever melhor os valores de Y; por outro lado, também precisaríamos de mais amostras para estimar bem $\mathbb{P}(X, Y)$, e quando isso não é possível erros de estimação fazem com que o custo de X suba.

Podemos então considerar um caso particular do problema de seleção de características em que a função de custo descreve "curvas em u" em todas as cadeias do reticulado booleano. Esse caso particular é conhecido como problema U-curve e existe na literatura algoritmos ótimos para esse problema como o `U-Curve Branch and Bound (UBB)`, `U-Curve-Search (UCS)` e `Poset Forest Search (PFS)`. A solução do problema U-Curve tem aplicações em problemas de aprendizado como projeto de W-operadores [JCB04] e preditores na estimação de Redes Gênicas Probabilísticas [Bar+07].

1.1 Objetivos do Trabalho

Capítulo 2

Conceitos Fundamentais

2.1 O problema de seleção de características

A seleção de características é um problema de otimização combinatória em que procuramos o melhor subconjunto de um conjunto de características S. O espaço de busca desse problema é o conjunto potência de S, $\mathcal{P}(S)$, que é uma coleção de todos os subconjuntos possíveis de S. A função de custo desse problema é uma função $c : \mathcal{P}(S) \rightarrow \mathbb{R}_+$.

Definição 2.1.1 (Problema de seleção de características) *Seja S um conjunto finito não vazio de características e c uma função de custo. Encontrar $X \in \mathcal{P}(S)$ tal que $c(X) \leq c(Y)$, $\forall Y \in \mathcal{P}(S)$.*

O espaço de busca do problema de seleção de características possui uma relação de ordem parcial definida pela relação $\subseteq$, portanto este conjunto é **parcialmente ordenado (poset)**.

Definição 2.1.2 *Uma **cadeia** do reticulado booleano é uma sequência X_1, X_2, ..., X_l tal que $X_1 \subset X_2 \subset \cdots \subset X_l$.*

Definição 2.1.3 *Uma cadeia é dita **maximal** se não existe outra cadeia no reticulado que contenha propriamente esta cadeia.*

Definição 2.1.4 *Uma função de custo c é dita decomponível em curvas u se para toda cadeia maximal $X_1, ..., X_l$, $c(X_j) \leq min\{c(X_i), c(X_k)\}$ sempre que $X_i \subset X_j \subset X_k$, $i, j, k \in \{1, ..., l\}$.*

2.2 Funções de custo para o problema de seleção de características

A função de custo utilizada na solução do problema de seleção de características deve, de alguma forma, refletir a qualidade do conjunto avaliado. Por isso, diferentes aplicações de seleção de características podem ter diferentes funções de custo. No contexto de aprendizado de máquina, uma função de custo utilizada é a entropia média condicional (MCE), utilizada por exemplo em projetos de W-operadores [DC 06].

Capítulo 3

Conclusão

Bibliografia

[Bar+07] J. Barrera, R. M. Cesar-Jr, D.C. Martins-Jr, R.Z.N Vencio, E. F. Merino, M. M. Yamamoto, F. G. Leonardi, C. A. B. Pereira e H. A. Portillo. «Constructing probabilistic genetic networks of *Plasmidium falciparum* from dynamical expression signals of the intraerythrocytic development cycle.» Em: *Methods of Microarray Data Analysis* V (2007), pp. 11–26.

[DC 06] R.M. Cesar-Jr an J. Barrera D.C. Martins-Jr. «W-operator window design by minimization of mean conditional entropy». Em: *Patter Analysis & Applications* (2006), pp. 139–153.

[JCB04] D. C. Martins Jr, R. M. Cesar-Jr e J. Barrera. «W-operator window design by maximization of training data information». Em: *Computer Graphics and Image Processing, 2004. Proceedings. 17th Brazilian Symposium* (2004), pp. 162–169.