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

Instituto de Matemática e Estatística

Gustavo Estrela de Matos

23 de Outubro de 2017

Resumo

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 a coleção de todos os subconjuntos possíveis de S. A função de custo desse problema é uma função $c : \mathcal{P}(S) \to \mathbb{R}_+$.

Definição 2.1.1 (Problema de seleção de características)**.** *Seja S um conjunto de características, finito e não vazio, 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 $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_l$.*

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 $X_1, ..., X_l$, $c(X_j) \leq max\{c(X_i), c(X_k)\}$ sempre que $X_i \subseteq X_j \subseteq X_k$, $i, j, k \in \{1, ..., l\}$.*

No contexto de aprendizado de máquina, é comum que as funções de custo utilizadas na seleção de característica descrevam curvas próximas do formato de u nas cadeias do reticulado. Esse fenômeno é conhecido em aprendizado e explicaremos como ele ocorre na seção 2.2.1.

2.2 Funções de custo

Nesta seção apresentaremos as duas funções de custo mais utilizadas durante este trabalho: a entropia condicional média (MCE) e a soma de subconjuntos. A primeira foi utilizada na seleção de modelos de aprendizado, enquanto a segunda foi utilizada para criação e solução de instâncias artificiais. A função de soma de subconjuntos é decomponível em curvas u e a MCE não, porém explicaremos como a última função deve ter um formato parecido com a da curva em u.

2.2.1 A curva em u no contexto de aprendizado

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

Definição 2.2.1. *Dado um problema de aprendizado em que Y é o conjunto de possíveis rótulos e $W = (w_1, ..., w_n)$, com $w_i \in A_i$, é o conjunto de variáveis. Seja $W' = (w_{I(1)}, w_{I(2)}, ..., w_{I(k)})$ um conjunto de variáveis (características) escolhidas, $\mathbf{X}$ uma vetor aleatório de tamanho k com $X_j \in A_{I(j)}$, e $log0 = 0$. Então, a **entropia condicional** de Y dado $\mathbf{X} = \mathbf{x}$ é:*

$$H(Y|\mathbf{X} = \mathbf{x}) = -\sum_{y \in Y} \mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}) log \mathbb{P}(Y = y|\mathbf{X} = \mathbf{x})$$

Definição 2.2.2. *Sob o mesmo contexto definido em 2.2.1, definimos a **entropia condicional média** como:*

$$\mathbb{E}[H(Y|\mathbf{X})] = \sum_{\mathbf{x} \in \mathbf{X}} H(Y|\mathbf{X} = \mathbf{x}) \mathbb{P}(\mathbf{X} = \mathbf{x})$$

A função H, em teoria da informação, mede o inverso da quantidade média de informação que uma variável tem. Esta função atinge valor máximo quando a distribuição de probabilidade da variável aleatória em questão é uniforme (todos valores que ela pode assumir são equiprováveis), e tem valores baixos quando essa distribuição é concentrada.

Problemas de aprendizado em que os rótulos tem uma distribuição concentrada são mais fáceis do que os problemas em que essa distribuição é menos concentrada. Tome como exemplo o problema de classificar o lançamento de uma moeda $\mathbf{x}$ em y (cara ou coroa); se toda moeda $\mathbf{x}$ é não viciada, então a distribuição de $\mathbb{P}(y|\mathbf{x})$ é pouco concentrada, por outro lado, quando a moeda é viciada, a distribuição de $\mathbb{P}(y|\mathbf{x})$ é concentrada e é mais fácil classificar este problema. Em termos mais formais, o erro do melhor classificador do problema mais fácil é menor do que o erro do melhor classificador do problema mais difícil.

Portanto, como a função H é capaz de medir a concentração da distribuição de Y dado $\mathbf{X} = \mathbf{x}$, e quanto maior esta concentração mais fácil é o modelo de aprendizado, podemos dizer que a função de custo $\mathbb{E}[H(Y|\mathbf{X} = \mathbf{x})]$ pode representar a qualidade do modelo de classificação que usa o conjunto de características de $\mathbf{X}$.

Agora, como já entendemos um pouco o funcionamento da função de custo MCE e como ela se relaciona com a qualidade do conjunto de características avaliado, vamos entender o que acontece no modelo de aprendizado e na função de custo que usamos como exemplo quando percorremos uma cadeia do reticulado.

Uma cadeia do poset pode ser vista como uma sequência de possíveis escolhas de conjuntos de características ao qual a cada passo adicionamos uma característica. Isso significa que a cada passo dado a variável $\mathbf{x}$ ganha uma componente a mais. Quando estamos no início da cadeia, poucas variáveis do problema são consideradas, portanto há uma grande abstração dos dados dos objetos sendo classificados, e conforme subimos uma cadeia, diminuímos a abstração dos dados e isso faz com que a distribuição de Y dado $\mathbf{x}$ se concentre.

Essa concentração da distribuição da probabilidade explica parcialmente a curva em u, porque indica apenas como o custo diminui conforme subimos uma cadeia do reticulado. Ao mesmo tempo, este raciocínio nos leva a pensar que adicionar características sempre melhora a classificação; de fato, o valor de $\mathbb{E}[H(Y|\mathbf{X} = \mathbf{x})]$ deve diminuir (até algum ponto de saturação)

conforme aumentamos o número de variáveis do problema. Mas se isso é verdade, por que fazemos seleção de características? A inconsistência entre esse raciocínio e a motivação para seleção de característica é que essa linha de raciocínio negligenciou que problemas de classificação (supervisada) dependem de uma amostra da distribuição de Y dado $\mathbf{X} = \mathbf{x}$, ou seja, não sabemos nem ao menos calcular $H(Y|\mathbf{X} = \mathbf{x})$, podemos apenas estimar.

A amostra da distribuição de Y dado $\mathbf{X} = \mathbf{x}$ é obtida do conjunto de treinamento do problema de aprendizado e quando o número de amostras não é grande o suficiente a qualidade do classificador é comprometida. Além disso, o número de amostras necessárias deve crescer conforme aumentamos a complexidade do modelo de aprendizado utilizado. Considerando que quando subimos uma cadeia do reticulado booleano estamos aumentando a complexidade do modelo, temos que, a partir de um certo ponto, a qualidade do classificador que utiliza tal conjunto de características deve piorar, o que justifica, junto a explicação anterior, a curva em u no custo das cadeias do reticulado.

No cálculo da entropia condicional média, o efeito do aumento da complexidade de $\mathbf{X}$ é a estimação ruim de $\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x})$. Contorna-se este problema modificando a entropia condicional média para penalizar a entropia de Y quando $\mathbf{x}$ foi observado poucas vezes. A função de custo utilizada é, então:

$$\hat{\mathbb{E}}[H(Y|\mathbf{X})] = \frac{N}{t} \sum_{\mathbf{x} \in \mathbf{X}} H(Y|\mathbf{X} = \mathbf{x}) \mathbb{P}(\mathbf{X} = \mathbf{x})$$

Note que esta função de custo não é necessariamente decomponível em curvas u, porém é provável que hajam poucas violações da regra nas cadeias do reticulado.

2.2.2 Soma de subconjuntos

Para se avaliar o desempenho dos algoritmos criados neste trabalho, utilizamos instâncias artificiais que são reduções do problema da soma de subconjuntos. Este problema consiste em, dado um conjunto finito de inteiros não-negativos S e um inteiro não-negativo t, descobrir se há um subconjunto de S que soma t. Podemos resolver este problema com a solução de uma instância do problema de seleção de características onde o conjunto de características é S' uma cópia de S e a função de custo é c:

$$c(X) = |t - \sum_{x \in X} x|, \text{ para todo } X \in \mathcal{P}(S').$$

Ao contrário da função de custo MCE, a função de custo de somas de subconjuntos é decomponível em curvas em u. Vamos provar esta propriedade agora.

Sejam A, B, C subconjuntos de uma cadeia do reticulado, tal que $A \subseteq B \subseteq C$. Vamos provar para dois casos disjuntos, quando $|t - \sum_{b \in B} b| > 0$ e quando $|t - \sum_{b \in B} b| \leq 0$. Começamos a demonstração definindo $D = B \setminus A$ e $E = C \setminus B$.

- se $|t - \sum_{b \in B} b| > 0$, então:

$$\begin{aligned} c(B) &= |t - \sum_{b \in B} b| \\ &\leq |t - \sum_{b \in B} b + \sum_{d \in D} d| \quad \text{(pois } S \text{ contém apenas números positivos e } t - \sum_{b \in B} b > 0) \\ &= |t - \sum_{a \in B \setminus D} a| \\ &= |t - \sum_{a \in A} a| \\ &= c(A) \end{aligned}$$

  portanto, $c(B) \leq c(A)$, logo $c(B) \leq max\{c(A), c(C)\}$.

- se $|t - \sum_{b \in B} b| \leq 0$, então:

$$\begin{aligned} c(B) &= |t - \sum_{b \in B} b| \\ &\leq |t - \sum_{b \in B} b - \sum_{e \in E} e| \quad \text{(pois } S \text{ contém apenas números positivos e } t - \sum_{b \in B} b \leq 0) \\ &= |t - \sum_{c \in B \cup E} c| \\ &= |t - \sum_{c \in C} c| \\ &= c(C) \end{aligned}$$

  portanto, $c(B) \leq c(C)$, logo $c(B) \leq max\{c(A), c(C)\}$.

Como acabamos de provar para os dois casos possíveis, temos que $c(B) \leq max\{c(A), c(C)\}$. $\square$

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.