Aprendizado Descritivo

Aula 14 – Mineração de modelos de sobrevivência excepcionais

Professor Renato Vimieiro

DCC/ICEx/UFMG

Introdução

- Imagine uma situação em que queremos identificar possíveis fatores de risco para uma doença
  - Identificar variáveis que possam afetar o prognóstico de pacientes
- Exemplificando, Milioli et al. (2017) investigaram a influência de fatores genéticos em subtipos de câncer de mama e como esses afetam o prognóstico (sobrevivência) das pacientes
- A intenção era identificar a existência de marcadores que subdividissem as pacientes em grupos mais homogêneos com prognósticos mais similares

Introdução

- Eles descobriram uma assinatura de 80 genes que subdividiam o tipo mais agressivo de câncer de mama em dois
- A figura ao lado mostra as curvas de sobrevivência para o tipo Basal (em cinza) como era considerado antes, e os novos grupos após a descoberta da assinatura
- O tipo representado pela linha em laranja (mais abaixo) é mais agressivo. A probabilidade de sobrevivência após 5 anos é de cerca de 60%

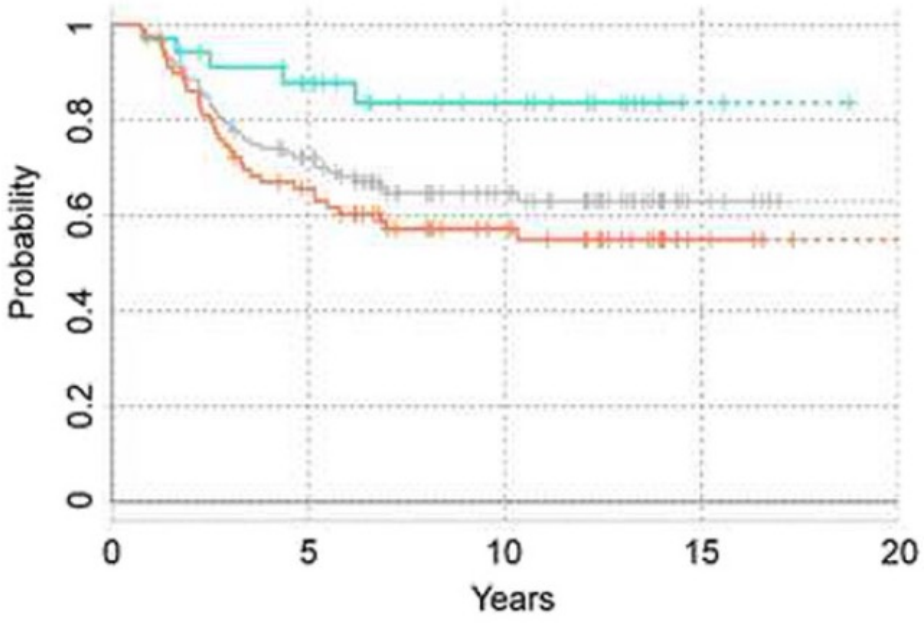


Figura retirada de Milioli et al. (2017)

Introdução

- Nessa aplicação, os autores propuseram uma abordagem para seleção desses 80 genes de forma semi-automatizada
  - Uma pré-seleção foi feita com base no conhecimento sobre o domínio do problema
  - Depois, uma análise univariada dos genes pré-selecionados foi feita para filtrar o conjunto final com base na diferença de sobrevivência de grupos induzidos entre baixa e alta expressão (valores menores que 1º ou maiores que 3º quartis)
- Note que a metodologia proposta por eles tenta responder à seguinte pergunta:
  - Existe algum subconjunto de genes associado a um grupo cuja curva de sobrevivência seja distinta (excepcional) em relação à população?

Introdução

- Essa pergunta pode ser reformulada para um contexto mais amplo, com dados de outras naturezas, se trocarmos a expressão ‘conjunto de genes’ por ‘conjunto de atributos’
- Em particular, se formos ainda mais além e substituirmos conjunto de atributos por conjunto de descrições, o problema levantado por eles é muito similar ao de descoberta de subgrupos ou ainda mineração de modelos excepcionais

Análise de sobrevivência

- A característica fundamental dos dados do exemplo anterior é que eles envolvem:
  - Um conjunto de atributos que descreve as amostras/objetos da base. No caso particular, os genes das pacientes estudadas.
  - Um atributo especial ‘tempo de sobrevivência’ que indicava o tempo total que a paciente sobreviveu desde o início do estudo
  - E um rótulo (atributo categórico), indicando se a paciente morreu em decorrência da doença ou não

Análise de sobrevivência

- Um exemplo do tipo de dados do problema seria como na tabela abaixo
- Cada objeto é uma linha da tabela
- Cada coluna é um gene (exceto a primeira e as duas últimas)
- A penúltima coluna indica se houve morte pela doença
- A última o tempo de sobrevivência

ID	EXO1	CENPF	NAT1	EGFR	FOXA1	δ	t
1	med	low	med	low	med	0	131
2	med	high	med	low	med	1	55
3	med	low	med	med	med	0	61
4	low	low	med	low	high	0	188
5	low	low	med	med	med	1	181
6	low	low	med	low	high	0	25
7	low	low	low	high	low	0	202
8	high	med	low	high	low	1	15
9	med	med	med	high	low	1	51

Análise de sobrevivência

- Nesse tipo de estudo, estamos interessados em construir modelos que sejam capazes de responder não só *SE* uma paciente irá falecer, mas qual a probabilidade de sobrevivência até um certo período de tempo
  - A área de estatística que lida com dados dessa natureza é conhecida como análise de sobrevivência
- A análise de sobrevivência estuda modelos para dados envolvendo o tempo até a ocorrência de um evento
  - Embora um tipo de evento recorrente nesses estudos seja a morte, ele pode ser a quebra de uma peça em uma máquina, a volta de uma doença, etc
  - O tempo de sobrevivência é tempo decorrido desde o início do estudo até a ocorrência do evento (ou perda de informação sobre o indivíduo – um paciente pode falecer por outras causas)

Análise de sobrevivência

- Repare que em análise de sobrevivência temos duas variáveis alvo, que são igualmente importantes:
  - Tempo
  - Evento
- Em um primeiro momento, poderíamos postular a construção de um modelo de sobrevivência através de regressão, já que estamos interessados em predizer o tempo de sobrevivência
  - Contudo, como o regressor não considera a informação do evento, muitos dados deveriam ser desprezados, pois os indivíduos não experenciam o evento
- Além disso, o fato de um indivíduo ter sobrevivido até um certo tempo, mas ter experienciado outro evento antes do término do estudo é relevante para o modelo de sobrevivência
  - Se uma paciente não faleceu em decorrência do câncer durante cinco anos de estudo, e depois tiver abandonado o estudo ou morrido por outras causas, ela deveria ser levada em consideração para ajustar a probabilidade de sobrevivência por até 5 anos
  - Seus dados deveriam ser desprezados somente após esse tempo

Análise de sobrevivência

- Dessa forma, o tempo de sobrevivência deve ser interpretado como o tempo sobrevivido enquanto tínhamos informação sobre o indivíduo
- Indivíduos que não experienciam o evento durante o estudo, mas, que por alguma razão, abandonaram o estudo são tratados como **dados censurados**
- Os modelos de sobrevivência devem levar em consideração todos os dados enquanto eles não forem censurados ou experienciarem o evento

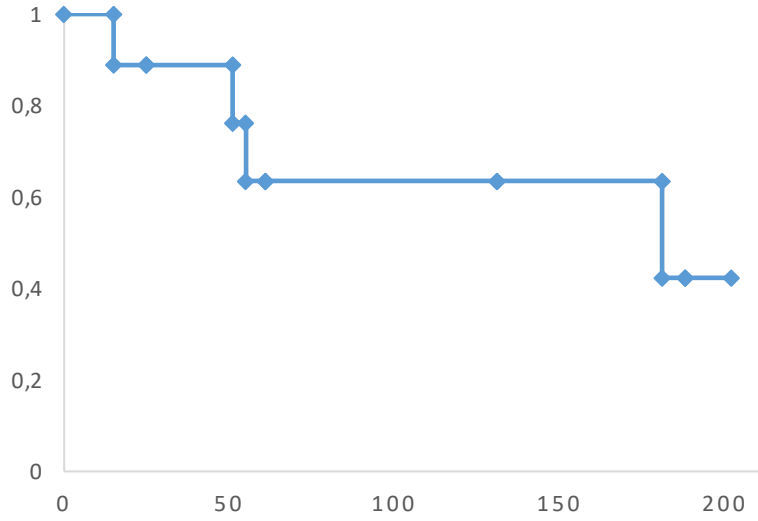
Análise de sobrevivência

- O modelo mais simples e amplamente usado na literatura médica é o modelo de Kaplan-Meier
- Eles definem que a probabilidade de sobrevivência até um tempo t_i é dado por:
  - $\hat{S}(t_i) = \hat{S}(t_{i-1})\left(1 - \frac{d_i}{r_i}\right)$
  - $\hat{S}(0) = 1$
  - d_i é o número de indivíduos que experienciaram o evento até o tempo t_i
  - r_i é o número de indivíduos que estavam em risco até o tempo t_i (não censurados e não experienciaram o evento)
- Em palavras, a estimativa de sobrevivência até o tempo t_i é a estimativa de sobreviver até o tempo anterior multiplicada pela probabilidade de não experienciar o evento no momento

Análise de sobrevivência

ID	EXO1	CENPF	NAT1	EGFR	FOXA1	δ	t
1	med	low	med	low	med	0	131
2	med	high	med	low	med	1	55
3	med	low	med	med	med	0	61
4	low	low	med	low	high	0	188
5	low	low	med	med	med	1	181
6	low	low	med	low	high	0	25
7	low	low	low	high	low	0	202
8	high	med	low	high	low	1	15
9	med	med	med	high	low	1	51

t_i	$\widehat{S}(t_i)$
0	1
15	0,88888889
25	0,88888889
51	0,76190476
55	0,63492063
61	0,63492063
131	0,63492063
181	0,42328042
188	0,42328042
202	0,42328042

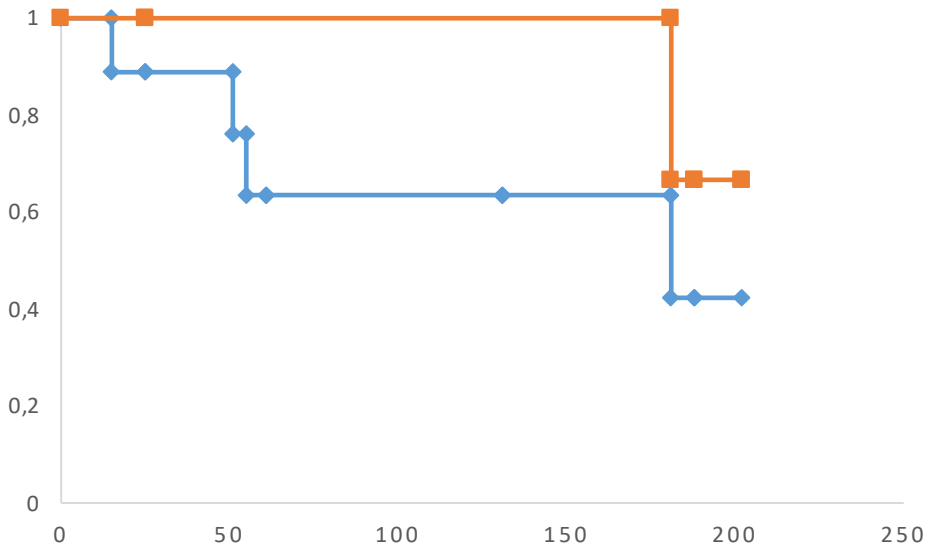


Análise de sobrevivência

- Em geral, os dados são estratificados de alguma forma para avaliar a relação de algum fator com a curva de sobrevivência
  - Avaliar se um tratamento é benéfico (melhor prognóstico) em relação a um placebo
- Nos dados do nosso exemplo, podemos avaliar a expressão de um gene está relacionada a um melhor prognóstico
- Por exemplo, EXO1='low' é relacionado a um melhor prognóstico?

Análise de sobrevivência

ID	EXO1	CENPF	NAT1	EGFR	FOXA1	Evento	t_i	$\widehat{S}(t_i)$
							0	1
6	low	low	med	low	high	0	25	1
5	low	low	med	med	med	1	181	0,66666667
4	low	low	med	low	high	0	188	0,66666667
7	low	low	low	high	low	0	202	0,66666667



Análise de sobrevivência

- Visualmente, percebemos que as curvas das amostras que possuem expressão EXO1=‘low’ e da população são bem distintas
  - O grupo possui um prognóstico melhor que a população no geral
- Contudo, validar estatisticamente essa diferença, podemos usar o teste de hipótese log-rank que verifica a hipótese nula de que não existe diferença nas distribuições em qualquer ponto no tempo
- Dessa forma, podemos usar o arcabouço de análise de sobrevivência para buscar e descrever subgrupos de amostras com um comportamento excepcional em relação à população/complemento

Exceptional Survival Model Mining (ESMAM)

- Mattos et al. (2020) apresentaram um algoritmo baseado em Otimização por Colônia de Formigas (ACO) para encontrar padrões em dados de sobrevivência
  - Exceptional Survival Model Ant Miner
- O objetivo central é fornecer uma ferramenta capaz de encontrar padrões relacionados a subgrupos com curvas de sobrevivência excepcionais
  - A ideia era fornecer aos especialistas do domínio uma alternativa mais automática para a proposta de Milioli et al. (2017)

ESMAM

- O ESMAM assume uma linguagem de descrição similar à do SD-Map
- São esperados somente dados categóricos
- Dados numéricos devem ser discretizados na etapa de pré-processamento
- Seletores são do tipo $a_i \in V_i$
  - Mas cada para $a_i = v_{ij}$ é tratado isoladamente

ESMAM

- O conjunto de seletores do nosso exemplo seria:
  - $I = \{EXO1 = low, EXO1 = med, EXO1 = high, CENPF = low, CENPF = med, CENPF = high, NAT1 = low, NAT1 = med, NAT1 = high, EGFR = low, EGFR = med, EGFR = high, FOXA1 = low, FOXA1 = med, FOXA1 = high\}$
- Uma descrição $d = \{EXO1 = low, EXO1 = med, FOXA1 = high\} \equiv EXO1 \in \{low, med\} \wedge FOXA1 = high$

ID	EXO1	CENPF	NAT1	EGFR	FOXA1	δ	t
1	med	low	med	low	med	0	131
2	med	high	med	low	med	1	55
3	med	low	med	med	med	0	61
4	low	low	med	low	high	0	188
5	low	low	med	med	med	1	181
6	low	low	med	low	high	0	25
7	low	low	low	high	low	0	202
8	high	med	low	high	low	1	15
9	med	med	med	high	low	1	51

ESMAM

- A medida de qualidade usada é 1- pvalue (log-rank)
- O algoritmo permite que o usuário escolha se a comparação será sobre o complemento ou população
- O algoritmo assume uma estratégia similar à do CN2-SD e tenta descobrir subgrupos não redundantes que maximizem a cobertura dos objetos na base
- Assim como no CN2-SD, existe um laço ‘externo’ com o objetivo de maximizar a cobertura, e um laço ‘interno’ com o objetivo de encontrar o melhor subgrupo
  - Na literatura de ACO, o laço interno é chamado de colônia e cada iteração é considerado a busca de uma formiga por alimento

ESMAM

$\mathbb{G} \leftarrow \emptyset, \quad G \leftarrow \emptyset$
$\mathbb{U} \leftarrow D, \quad \Delta U \leftarrow 0$
$stag \leftarrow 0$
while $\mathbb{U} \neq \emptyset$ ***and*** $stag \leq maxStag$ **do**
    $\texttt{searchInitialization}(I(G), \mathbb{G}, \mathbb{I}, \mathbb{U}, \wp^H)$
    $G \leftarrow \texttt{subgroupSearch}(\mathcal{B}, \wp^S)$
    $\mathbb{G} \leftarrow \texttt{subgroupSetUpdating}(G, \mathbb{G}, \alpha)$
    $\Delta U \leftarrow |\mathbb{U}| - |\bigcup \mathbb{G}|$
    $\mathbb{U} \leftarrow \bigcup G$
    if $\Delta U = 0$ **then**
        $stag \leftarrow stag + 1$
    else: $stag \leftarrow 0$
return: $\mathbb{G}$

ESMAM

- Como a metáfora é a busca de uma colônia de formigas por alimento, assim como na natureza, as formigas iniciam a busca de forma ‘aleatória’ e, ao encontrarem indícios de comida, sinalizam às outras através de feromônios deixados no caminho
- As próximas formigas que saírem da colônia tenderão a seguir esse caminho promissor, evitando buscas aleatórias, e, caso encontrem mais indícios de comida, reforçam o feromônio deixado no caminho, influenciando mais formigas a seguirem por ele
- Transportando essa ideia para o nosso contexto, o caminho a ser explorado pelas formigas é o espaço de busca de descrições
- Caminhos promissores são aqueles que incluem seletores que gerem subgrupos mais excepcionais com respeito à medida de qualidade
  - Se uma formiga percebe que um seletor é bom, ela sinaliza esse fato aumentando o feromônio relacionado a ele

ESMAM

- Inicialmente, o feromônio dos seletores é tido como uniforme
- O feromônio de um seletor $(a_i = v_{ij}) = I_{ij} \in I$ é:
  - $\tau_{ij} = 1/|I|$
- Isso é consistente com a ideia de que, inicialmente, as formigas executam buscas ‘aleatórias’ por comida, já que a escolha de um seletor para compor uma descrição segue a distribuição definida pelo feromônio
- À medida que descrições são encontradas, o feromônio dos seletores é atualizado para refletir a qualidade das escolhas
  - $\tau_{ij}^{q+1} = (1 + \varphi(d)) \cdot \tau_{ij}^{q}$
- Uma etapa de normalização é feita para ajustar o valor do feromônio para o intervalo desejado

ESMAM

- Como forma de auxiliar no processo de busca, podemos ‘induzir’ as formigas para direções que achamos mais promissoras
- Essa ‘indução’ é feita através de uma função heurística que enviesa a busca na direção de seletores mais promissores
  - Ela pode, por exemplo, atribuir uma maior probabilidade de escolha a seletores mais úteis do ponto de vista do domínio do problema
  - Por exemplo, podemos aumentar a probabilidade de seletores envolvendo genes sabidamente relacionados à doença serem escolhidos

ESMAM

- A função heurística é definida no ESMAM da seguinte forma:
  - $\eta(I_{ij}) = \eta_H(I_{ij}) \cdot \eta_L(I_{ij}) \cdot \eta_W(I_{ij})$
- Os três componentes refletem:
  - A qualidade do seletor em discriminar os grupos
  - A frequência do seu uso nas descrições encontradas
  - A cobertura de objetos ainda não cobertos por outros subgrupos
- A qualidade do seletor é medida em termos de entropia com respeito ao tempo de sobrevivência médio
  - Se um seletor filtrar de forma homogênea objetos com tempo de sobrevivência acima/abaixo da média, então ele é mais interessante

ESMAM

ID	EXO1	CENPF	NAT1	EGFR	FOXA1	δ	t
1	med	low	med	low	med	0	131
2	med	high	med	low	med	1	55
3	med	low	med	med	med	0	61
4	low	low	med	low	high	0	188
5	low	low	med	med	med	1	181
6	low	low	med	low	high	0	25
7	low	low	low	high	low	0	202
8	high	med	low	high	low	1	15
9	med	med	med	high	low	1	51

t≥μ	H
0,99	0,08079314
0,95	0,28639696
0,9	0,46899559
0,8	0,72192809
0,7	0,8812909
0,6	0,97095059
0,5	1

$$H(t \geq 101|EXO1 = low) = -\left(\frac{1}{4} * \log\left(\frac{1}{4}\right) + \frac{3}{4} * \log\left(\frac{3}{4}\right)\right) = 0{,}81$$

ESMAM

- O componente de frequência de uso penaliza seletores que sejam escolhidos com frequência
- Ele é definido por:
- $\eta_L(I_{ij}) = 1 - \frac{1}{1+e^{-(s(I_{ij})-L)}}$
- O parâmetro L determina o número de vezes em que o uso do seletor é penalizado em 50%

ESMAM

- Finalmente, o componente de cobertura utiliza um esquema de pesos na cobertura tal como no CN2-SD
- $\eta_W(I_{ij}) = \left(\frac{1}{|c(I_{ij})|}\right) \sum_{o \in c(I_{ij})} W^{g(o)}$
- $g(o)$ é o número de subgrupos já encontrados que cobrem o objeto o
- $W \in (0,1]$ é um parâmetro definido pelo usuário
  - Mais próximo de 1, menor penalização

ESMAM

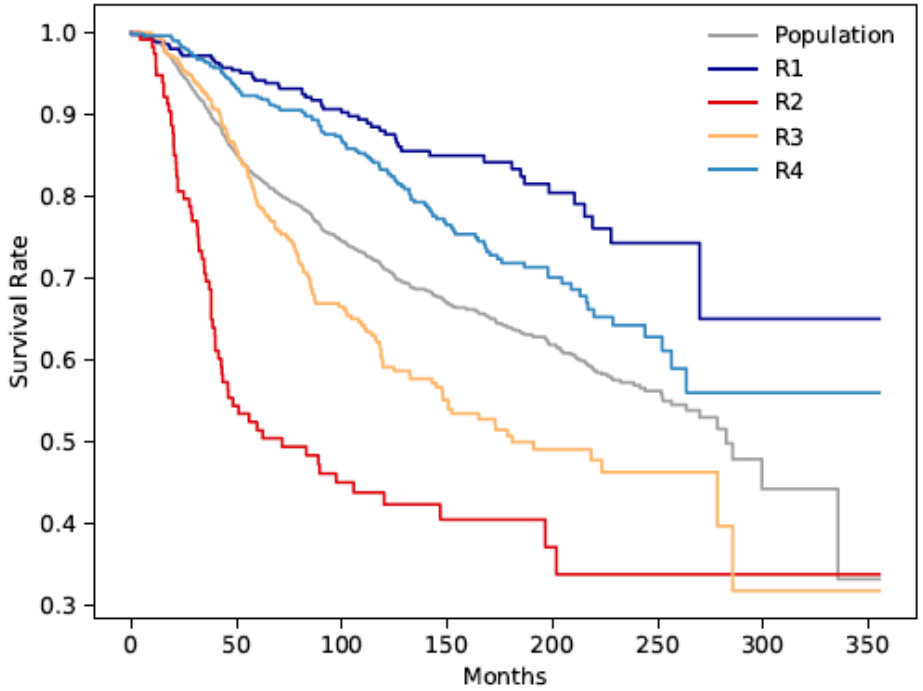
- Dessa forma, as formigas escolhem os seletores com a seguinte probabilidade
- $P(I_{ij}) = \frac{\iota(a_i)\cdot\eta(I_{ij})\cdot\tau_{ij}}{\sum \iota(a_i)\cdot\eta(I_{ij})\cdot\tau_{ij}}$
- A função $\iota(a_i)$ é uma função indicadora se o atributo já foi usado em algum seletor da descrição
- Somente os seletores que cobrem os mesmos objetos que a descrição atual cobre são considerados

ESMAM

- A busca pelos melhores subgrupos então é feita da seguinte forma

```
1  Function subgroupSearch(B̄, nAnts, nConverg, minCov):
2      t ← 0; converg ← 0
3      G⁻ ← ∅; G^best ← ∅
4      while t ≤ nAnts and converg ≤ nConverg do
5          G ← buildDescription(minCov)
6          G ← pruneDescription(G, B̄)
7          pheromoneUpdating(G)
8          if φ(G, B̄) > φ(G^best, B̄) then
9              G^best ← G
10         if G = G⁻ then
11             converg ← converg + 1
12         else: converg ← 0
13         G⁻ ← G
14         t ← t + 1
15     return: G^best
```

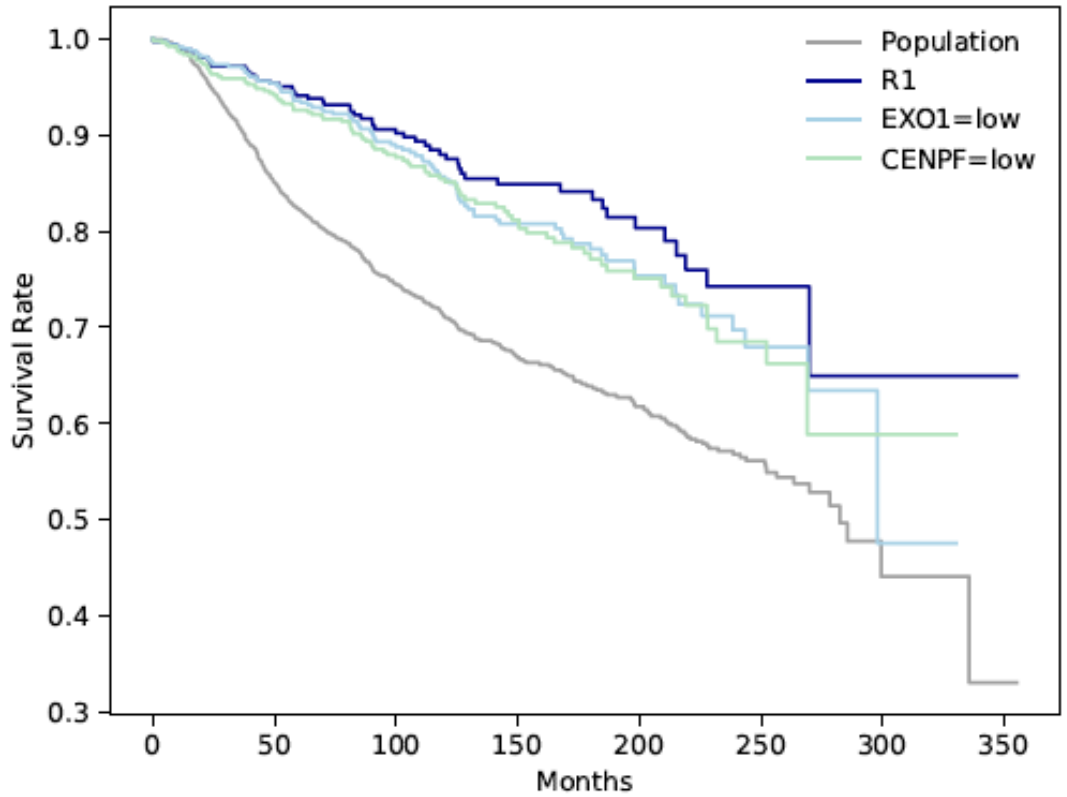
Aplicação em um conjunto de câncer de mama



ID	subgroup description	size	avg surv	median surv
–	Population	1980	125.22	116.45
R1	CENPF=low and EXO1=low	357	131.88	124.10
R2	GRB7=high and EGFR=high and ESR1=low	115	81.57	46.17
R3	CEP55=med and NAT1=med and MKI67=med	353	112.93	101.27
R4	NAT1=high	495	138.90	132.10

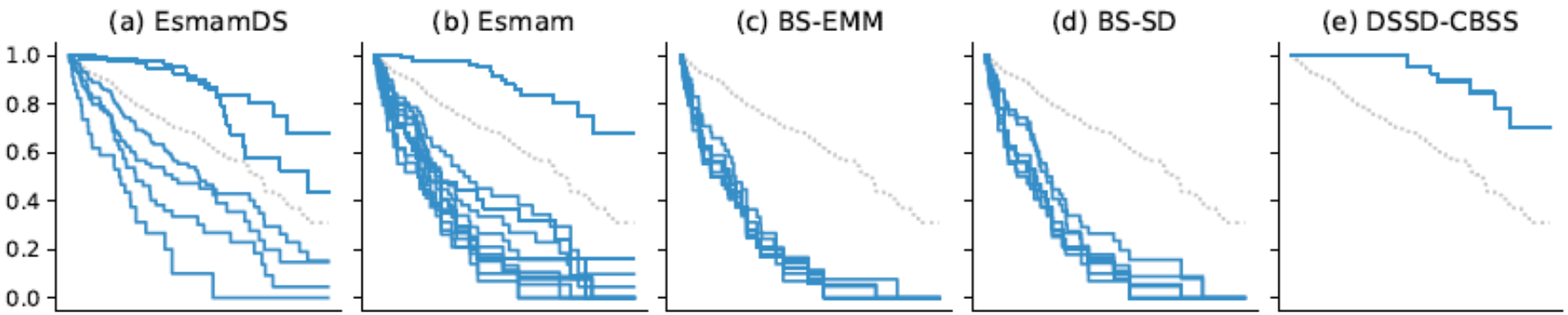
	Population	R1	R2	R3	R4
LumA	35.0	60.0	1.0	41.0	65.0
LumB	24.0	3.0	–	30.0	29.0
Her2	11.0	1.0	46.0	11.0	1.0
Claudin-low	11.0	19.0	8.0	7.0	1.0
Basal	11.0	–	39.0	2.0	–
Normal	7.0	17.0	6.0	9.0	3.0

Aplicação em um conjunto de câncer de mama

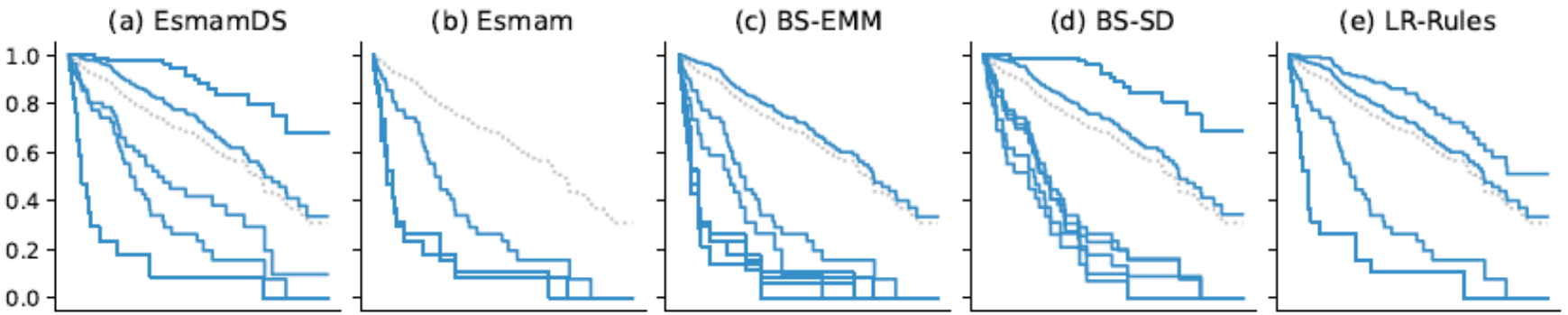


Comparação com SD

População



Complemento



Leitura

- Milioli, H.H., Tishchenko, I., Riveros, C. et al. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. BMC Med Genomics 10, 19 (2017). https://doi.org/10.1186/s12920-017-0250-9
- Mattos, J.B., Silva, E.G., de Mattos Neto, P.S.G., Vimieiro, R. (2020). Exceptional Survival Model Mining. In: Cerri, R., Prati, R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science(), vol 12320. Springer, Cham. https://doi.org/10.1007/978-3-030-61380-8_21
- Vimieiro, R., Mattos, J.B, de Mattos Neto, P.S.G. (2024). EsmamDS: A more diverse exceptional model mining approach. In: Information Sciences (under review).

Aprendizado Descritivo

Aula 14 – Mineração de modelos de sobrevivência excepcionais

Professor Renato Vimieiro

DCC/ICEx/UFMG