

Árvore de Decisão



George Darmiton da Cunha Cavalcanti
Tsang Ing Ren
CIn/UFPE

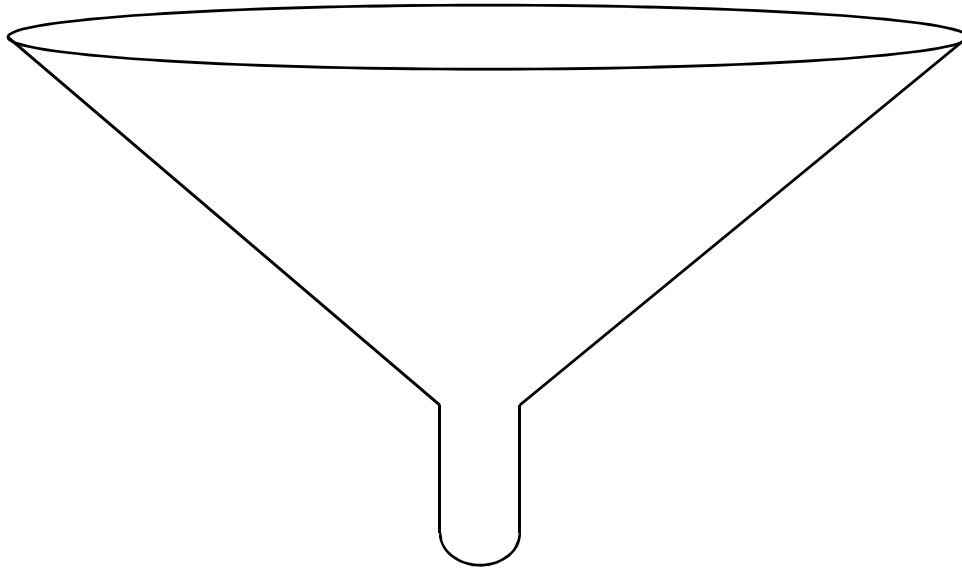
Tópicos

- Introdução
- Representando Árvores de Decisão
- O algoritmo ID3
- Definições
 - Entropia
 - Ganho de Informação
- *Overfitting*

Objetivo da aprendizagem

Conhecimento em extensão

(exemplos percepção-ação,
características-conceitos, etc.)



Conhecimento em intenção

(regras definições.)

Exemplos

dia 29, a Caxangá estava
engarrafada

dia 30, a Caxangá estava
engarrafada

dia 01, a Caxangá estava
engarrafada

dia 03, a Caxangá estava
engarrafada

Hipótese indutiva

Todo dia, a Caxangá está
engarrafada

Aprendizagem indutiva

- Inferência de uma regra geral (hipótese) a partir de exemplos particulares

- Exemplo: trânsito na caxangá

- Relação:

Precisão

versus

Quantidade de exemplos

Aprendizagem indutiva

□ Categorias

■ **Incremental**

- atualiza hipótese a cada novo exemplo
- mais flexível
- porém a ordem de apresentação é importante

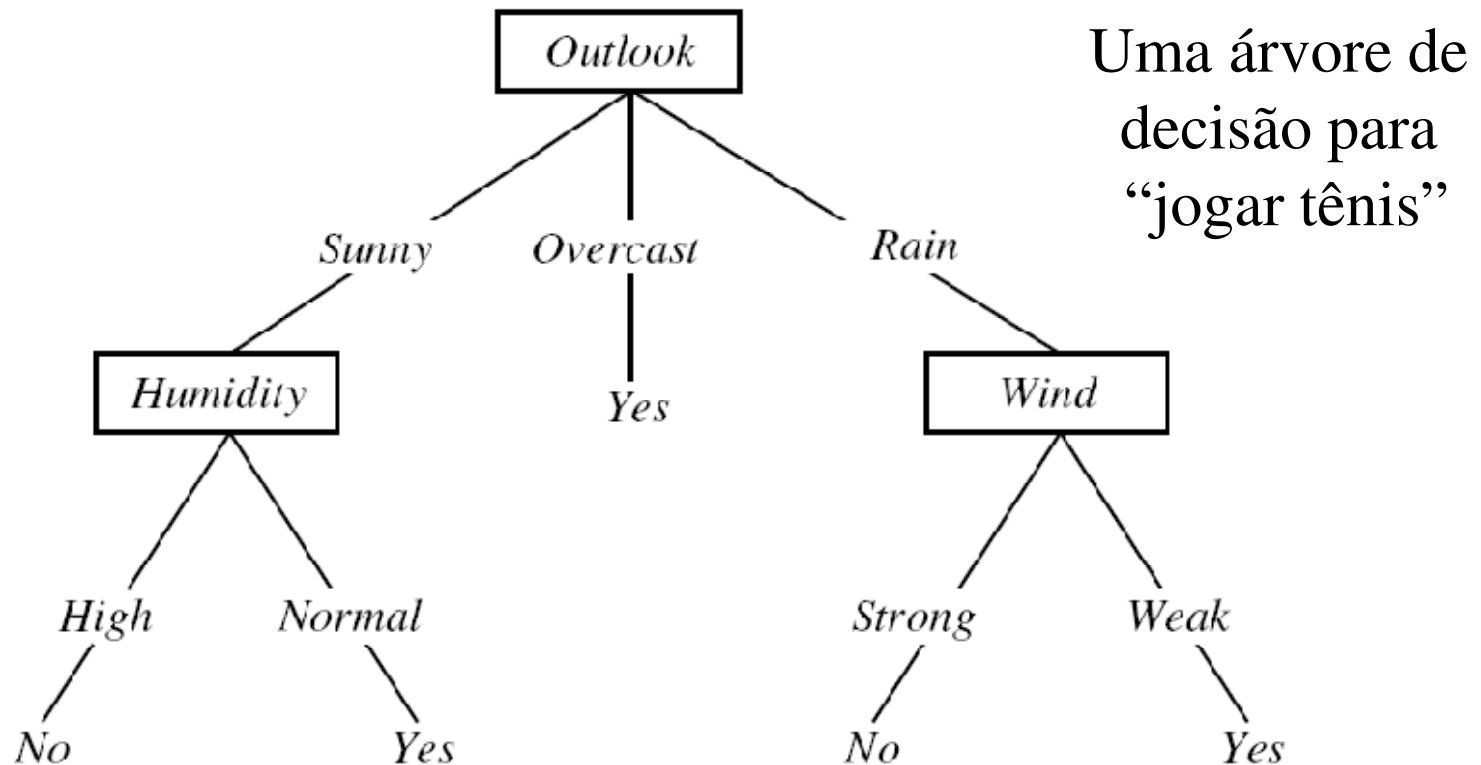
■ **Não Incremental**

- gerada a partir de todo conjunto de exemplos
- mais eficiente e prática

Uma Abordagem típicas em aprendizagem simbólica

- Árvores de decisão: *inductive decision trees* (ID3)
 - Lógica de ordem 0
 - Instâncias (exemplos) são representadas por pares atributo-valor
 - Fáceis de serem implementadas e utilizadas
 - Aprendizagem não incremental

Representando Árvores de Decisão



O que acontece se:

<Outlook=Sunny, Temperature=Hot, Humidity=High, Wind=Weak>?

Representando Árvores de Decisão

□ Árvore de decisão

- Cada nó interno testa um atributo
- Cada ramo corresponde a um valor do atributo
- Cada folha representa uma classe

Problemas apropriados para serem resolvidos através de Árvores de Decisão

- Instâncias são representadas por pares atributo-valor
 - Exemplos:
 - *Temperature* \leftarrow (*Hot, Mild, Cold*)
 - *Temperature* \leftarrow um número real
- A função objetivo possui valores discretos
 - O exemplo previamente apresentado possui duas saídas possíveis: *yes* e *no*
- Descrições disjuntivas são requeridas
 - Em geral, uma árvore de decisão representa uma disjunção de conjunções

Problemas apropriados para serem resolvidos através de Árvores de Decisão

- ❑ O conjunto de treinamento pode conter erros
 - Árvore de decisão é robusta a erros tanto em padrões do conjunto de treinamento quanto valores de atributos
- ❑ O conjunto de treinamento pode não possuir valores para alguns atributos
 - Árvores de decisão podem ser usadas mesmo na presença de valores desconhecidos

Problemas apropriados para serem resolvidos através de Árvores de Decisão

- Alguns problemas possuem as características apresentadas
 - Diagnóstico de doenças
 - Diagnóstico de mal funcionamento de equipamentos
 - Análise de crédito

Indução *Top-Down* em Árvores de Decisão

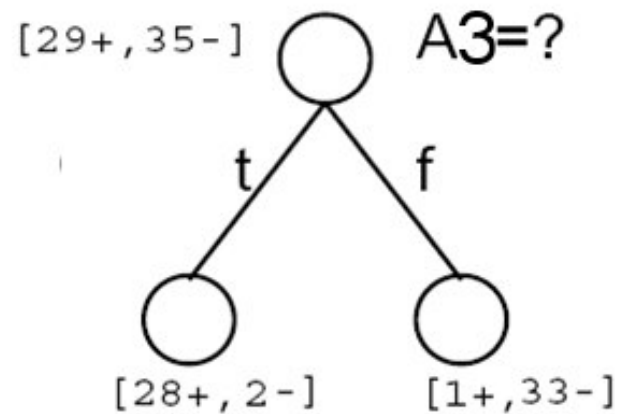
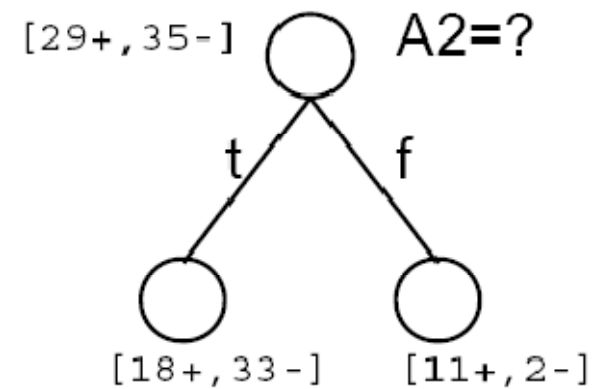
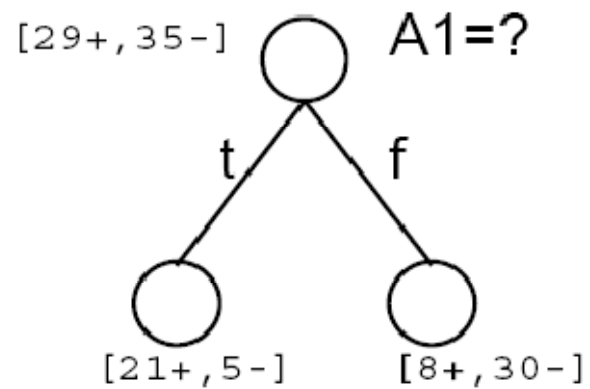
□ Laço principal

- $A \leftarrow$ o “melhor” atributo para o próximo nó
- A é atribuído ao nó
- Para cada valor de A, crie um novo ramo do nó
- Ordene os exemplos de treinamento para as folhas
- Se os padrões de treinamento são perfeitamente classificados Então PARE, Senão repita sobre novos nós

Critérios para Escolha do Atributo

- Como medir a *habilidade* de um dado atributo na tarefa de discriminar as classes?
- Existem muitas medidas.
- Todas concordam em dois pontos:
 - Uma divisão que mantém as proporções de classes em todas as partições é inútil;
 - Uma divisão na qual em cada partição todos os exemplos são da mesma classe tem utilidade máxima.

Qual é o melhor atributo?



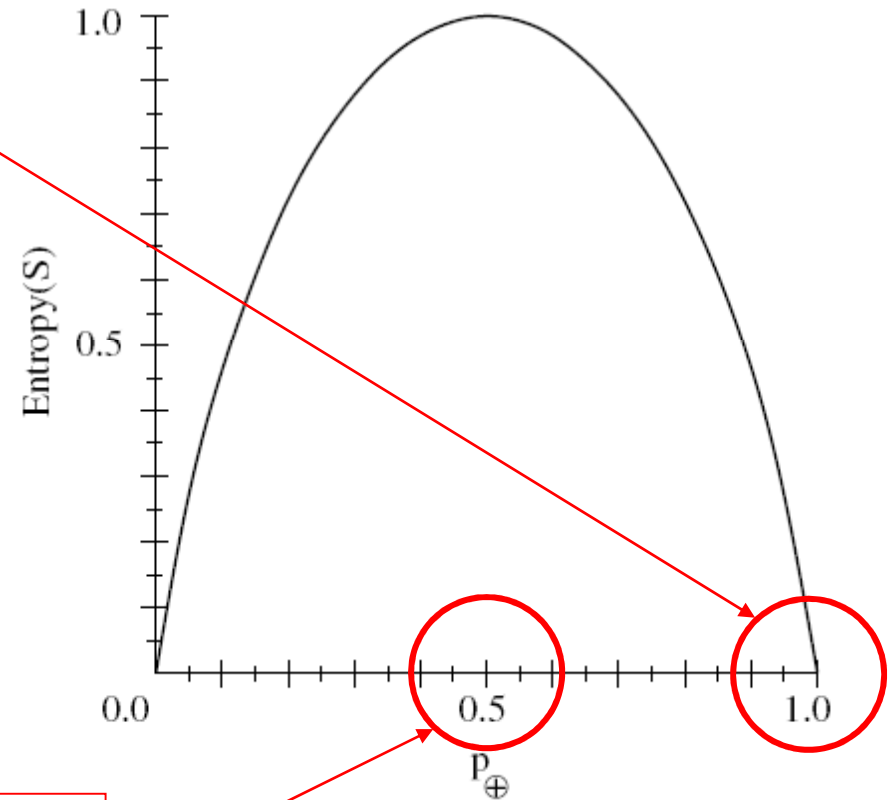
Entropia

- S é uma amostra dos exemplos de treinamento
- p_{\oplus} é a proporção de exemplos positivos em S
- p_{\ominus} é a proporção de exemplos negativos em S
- Entropia mede a *impureza* de S :

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Entropia

Se $p_{\oplus}=1$, exemplo positivo.
Nenhuma mensagem precisa ser enviada.
Entropia é 0 (mínima).



Se $p_{\oplus}=0.5$, um bit é necessário para
indicar se o exemplo é \oplus ou \ominus .
Entropia é 1 (máxima).

Entropia: Exemplo

- Suponha que S é uma coleção de 14 exemplos, incluindo 9 positivos e 5 negativos
 - Notação: $[9+,5-]$
- A entropia de S em relação a esta classificação booleana é dada por:

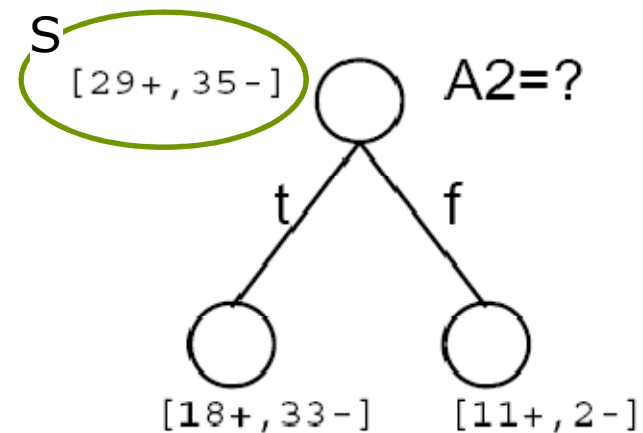
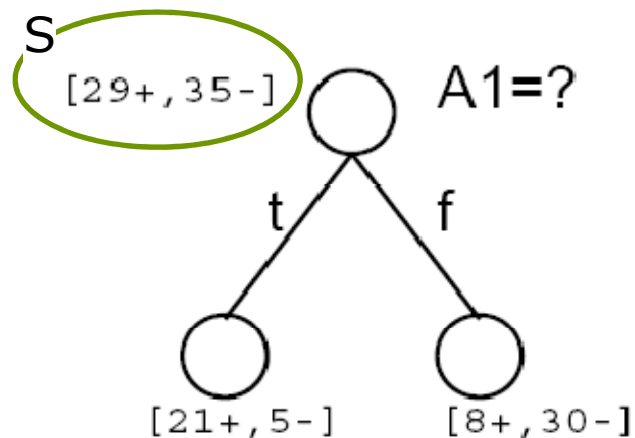
$$\begin{aligned} \text{Entropy}([9+,5-]) &= -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) \\ &= 0.940 \end{aligned}$$

Ganho de Informação

- Dado um conjunto de exemplos, qual atributo escolher?
 - Os valores de um atributo definem partições do conjunto de exemplos.
 - O ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo.

Ganho de Informação

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

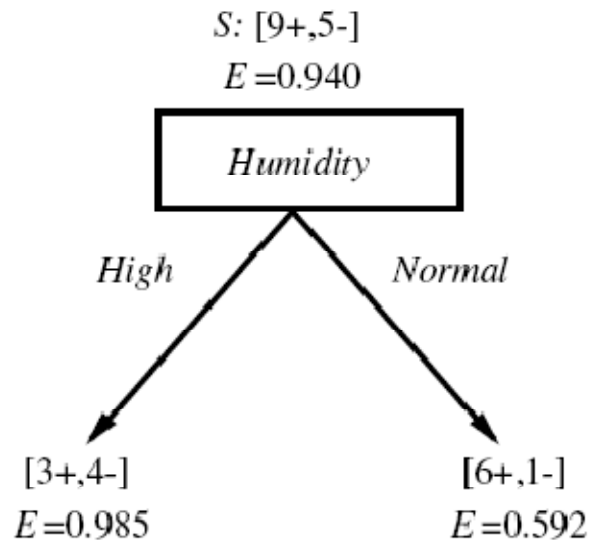


Padrões de Treinamento

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Selecionando o próximo atributo

Qual atributo é o melhor classificador?



$Gain(S, Humidity)$

$$= .940 - (7/14) \cdot .985 - (7/14) \cdot .592$$
$$= .151$$

Construção de uma Árvore de Decisão

Exemplos de Treino					
Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Referência : <http://bioinformatics.ath.cx>

Construção de uma Árvore de Decisão



Construção de uma Árvore de Decisão



Construção de uma Árvore de Decisão

Escolher o melhor atributo:

?

$$\text{Entropia}(S) = -p_+ \cdot \log_2 p_+ - p_- \cdot \log_2 p_-$$

$$\text{Ganho}(S,A) = \text{Entropia}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot \text{Entropia}(S_v)$$

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Tênis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Construção de uma Árvore de Decisão

$S = [9+, 5-]$
 $E = 0.940$

$\text{MAX} \left(\begin{array}{l} \text{Ganho}(S, \text{Humidade}) = 0.151 \\ \text{Ganho}(S, \text{Vento}) = 0.048 \\ \text{Ganho}(S, \text{Aspecto}) = 0.247 \end{array} \right) =$
 $= \text{Ganho}(S, \text{Aspecto})$

Aspecto
 Sol Nuvens Chuva

$[2+, 3-]$ $[4+, 0-]$ $[3+, 2-]$
 $E=0.971$ $E=0$ $E=0.971$

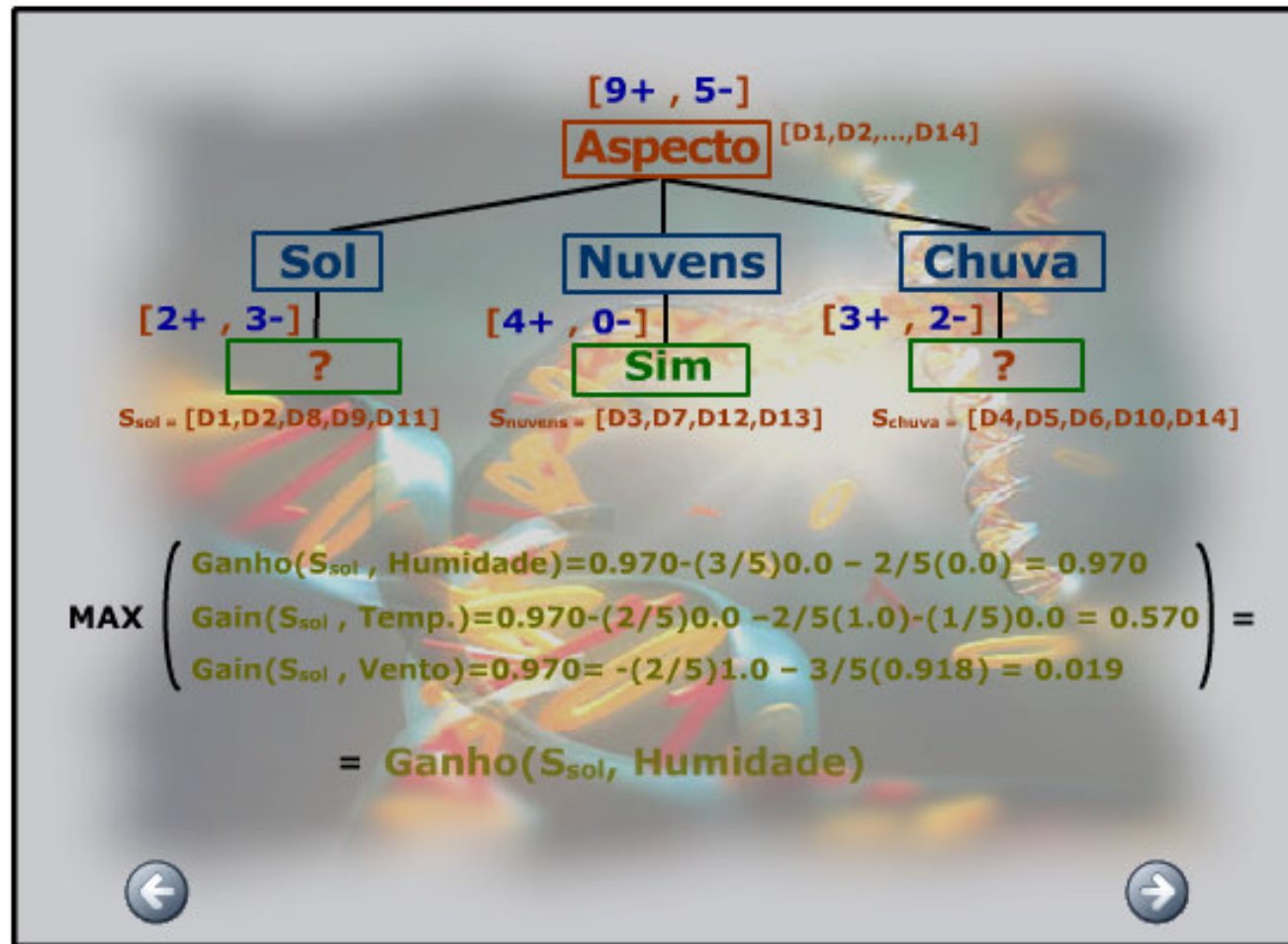
Ganho(S, Aspecto) =
 $= 0.940 - (5/14) \cdot 0.971$
 $- (4/14) \cdot 0.0$
 $- (5/14) \cdot 0.971$
 $= 0.247$

Entropia(S) = $-p_+ \cdot \log_2 p_+ - p_- \cdot \log_2 p_-$

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Tênis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

$\text{Ganho}(S, A) = \text{Entropia}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot \text{Entropia}(S_v)$

Construção de uma Árvore de Decisão



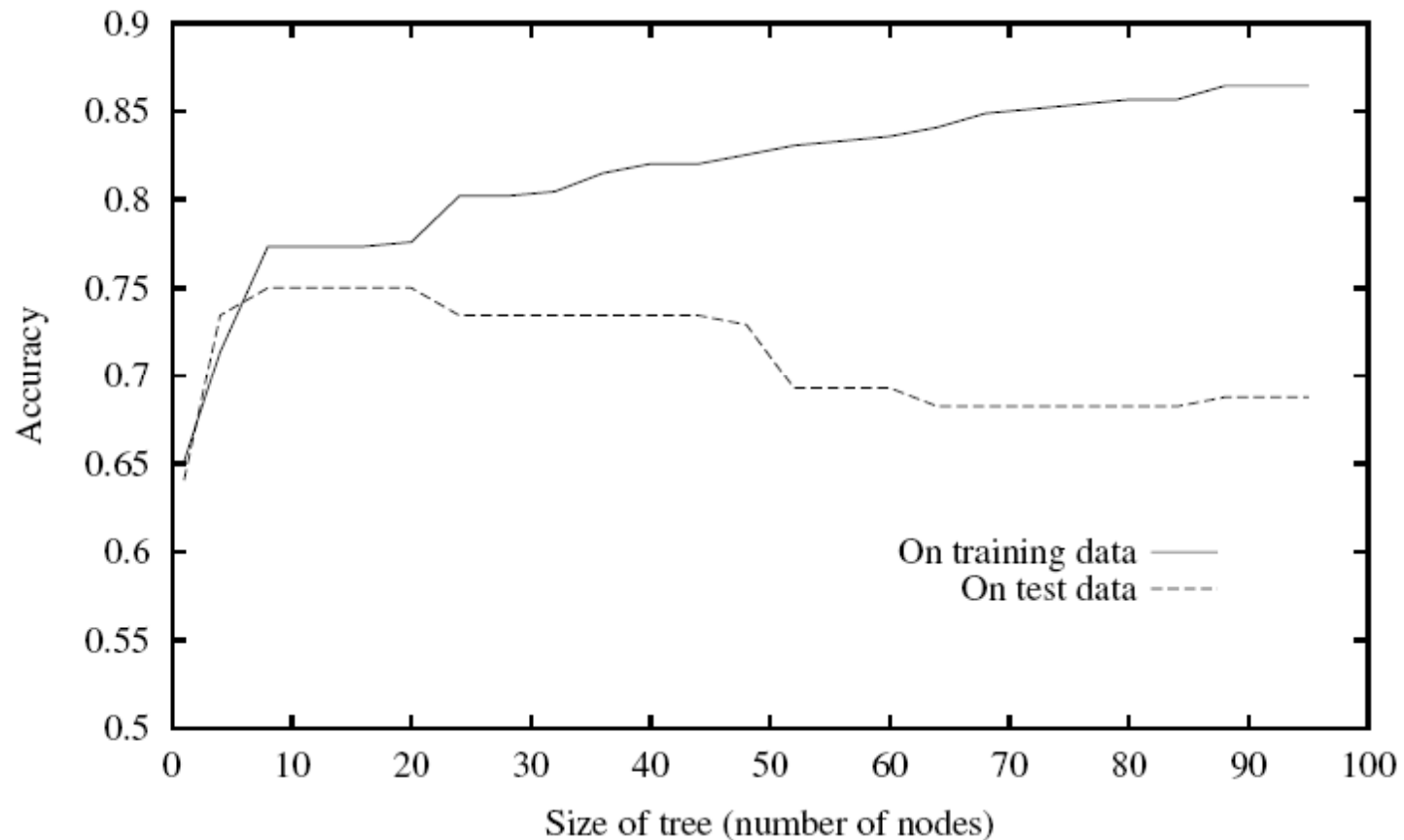
Construção de uma Árvore de Decisão



Espaço de Busca do ID3

- ❑ Sem *backtracking*
 - Mínimo local
- ❑ Escolhas de busca baseada em estatística
 - Robusta a dados ruidosos
- ❑ Preferência por árvores menores
 - Navalha de Occam

Overfitting



Pode ocorrer quando os dados possuem ruído ou quando o número de exemplos de treinamento é pequeno

Como evitar *overfitting*?

□ Alternativas

- Parar o crescimento antes que a árvore classifique os dados de treinamento perfeitamente
 - Permitir o completo crescimento da árvore e podá-la
-
- A primeira alternativa parece ser mais direta
 - Embora, a segunda tenha encontrado melhores resultados na prática
 - Pois, é difícil estimar precisamente o momento de parar

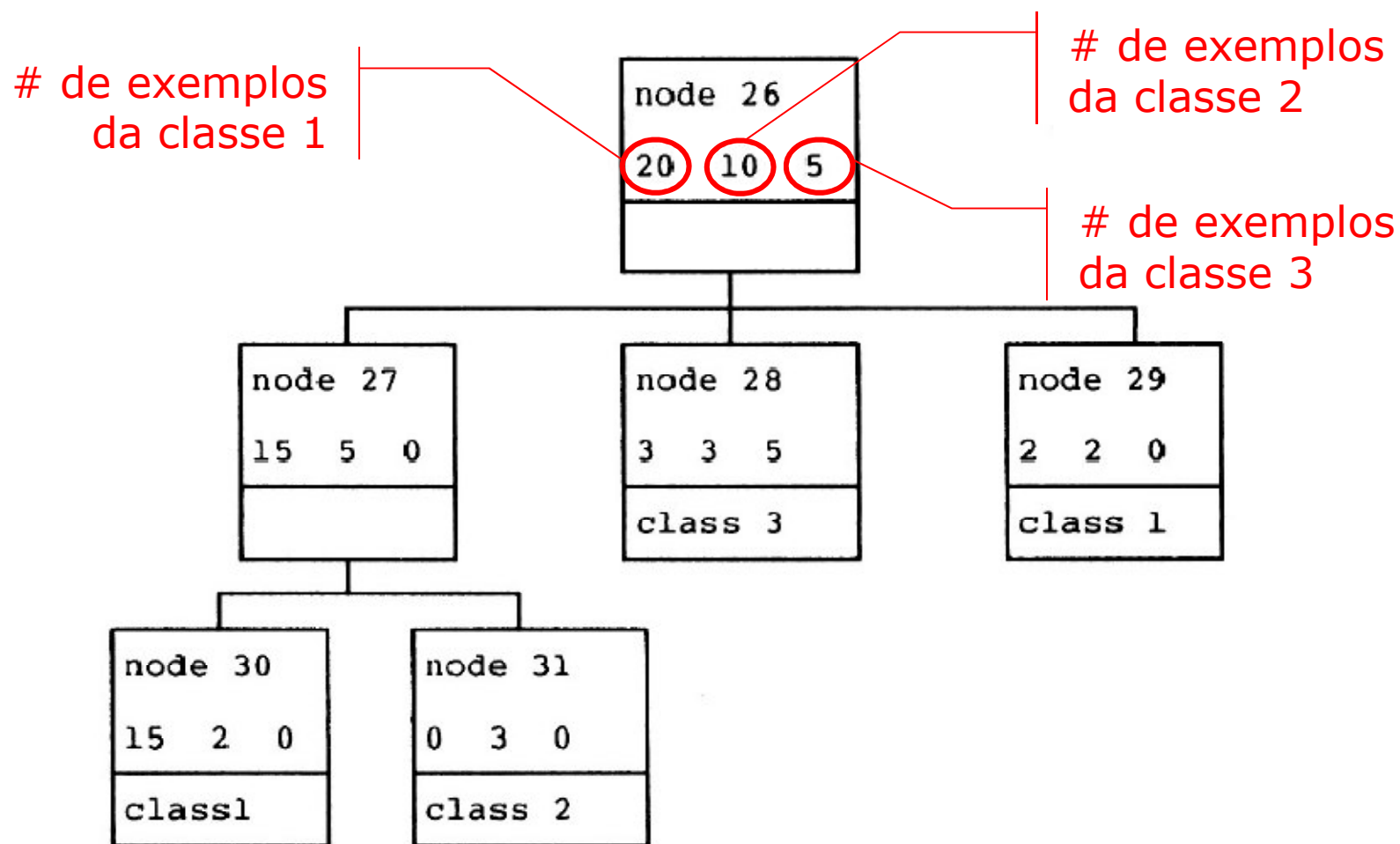
Como selecionar a melhor árvore?

- Independente da alternativa usada, como escolher a melhor árvore?
 - Calculando o desempenho sobre o conjunto de dados de treinamento
 - Calculando o desempenho sobre um conjunto de validação
 - MDL (*Minimum Description Length*)
 - minimize ($size(tree) + size(misclassifications(tree))$)

Abordagens para podar árvores

- Três estratégias
 - *Error-Complexity Pruning*
 - *Critical Value Pruning*
 - *Reduced-Error Pruning*

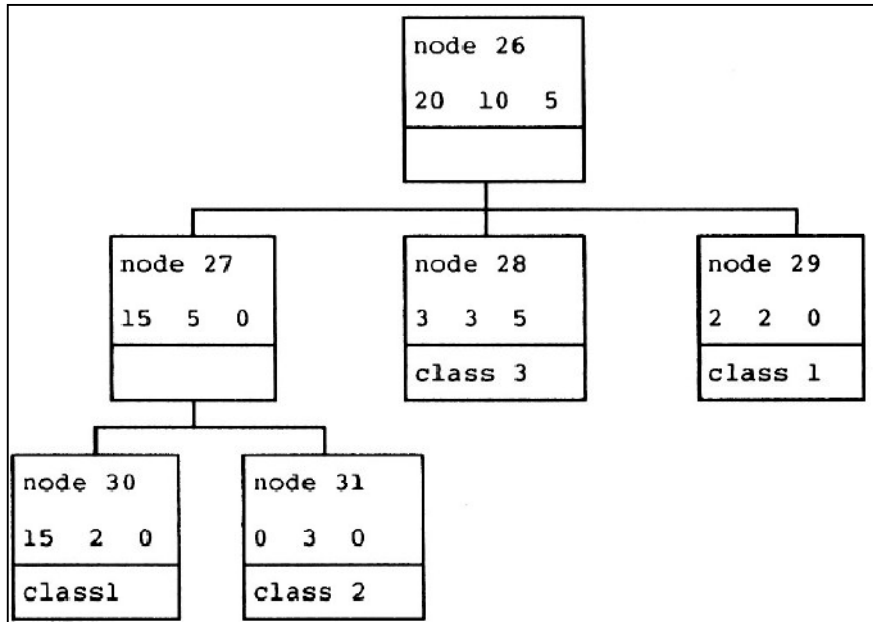
Exemplo de uma árvore parcialmente podada



Error-Complexity Pruning

- O método funciona da seguinte forma:
 - Cada nó é um ponto de partida para uma sub-árvore
 - Antes da poda, as folhas contém exemplos que pertencem a apenas uma classe
 - Após a poda, a folha conterá exemplos de diversas classes
 - Assim, a classe dessa folha é dada pela classe com maior frequência dentre os exemplos
 - Isso gera erro
 - Dividindo esse erro pelo número de folhas obtém-se uma medida de redução do erro por folha
 - Essa é a medida *error-complexity*

Error-Complexity Pruning (exemplo)



- Total de 200 exemplos
- t é um nó
- T é a sub-árvore
- Observando o nó 26
 - Possui 4 folhas, $N_T=4$
 - Caso esse nó seja podado
 - Ele será da classe 1
 - Assim, 15 dos 35 exemplos serão incorretamente classificados

- Assim, $r(t)=15/35$ e a proporção dos dados em t é $p(t)=35/200$.
- O custo do nó t é:

$$R(t) = r(t)p(t) = \frac{15}{35} \times \frac{35}{200} = \frac{15}{200}$$

Error-Complexity Pruning (exemplo)

- Caso a árvore não fosse podada, o custo da sub-árvore seria:

$$\begin{aligned} R(T_l) &= \sum R(i), \quad \text{for } i = \text{sub-tree leaves} \\ &= \frac{2}{17} \times \frac{17}{200} + 0 + \frac{6}{11} \times \frac{11}{200} + \frac{2}{4} \times \frac{4}{200} = \frac{10}{200} \end{aligned}$$

O algoritmo calcula α para cada sub-árvore e escolhe a que contém o menor valor para podar.

- Assim, o custo total da sub-árvore é $= R(T_l) + \alpha N_T$
- E quando a sub-árvore for podada $= R(t) + \alpha$
- Igualando as equações

$$\alpha = \frac{R(t) - R(T_l)}{N_T - 1} = \frac{15/200 - 10/200}{4 - 1} = 5/600$$

Critical Value Pruning

- ❑ Esse método observa para valores que medem a importância de cada nó
- ❑ Essa medida é calculada na criação da árvore
- ❑ Esse valor mostra quão bem o atributo divide os dados

Esse método especifica um valor crítico e poda os nós que não atingem o referido valor, a menos que um nó mais profundo não o atinja também

Critical Value Pruning

- Quanto maior o valor crítico, maior o grau de poda da árvore e menor a árvore resultante
- Na prática, uma série de árvores são geradas aumentando o valor crítico

Reduced-Error Pruning

- ❑ O método funciona da seguinte forma:
 - Comece com um árvore completa
 - Apresente dados de validação para a árvore
 - Para cada nó que não seja folha
 - ❑ Conte o número de erros caso a poda seja realizada e caso não seja
 - A diferença entre esses valores (se positiva) é uma medida de ganho do processo de poda
 - Esse processo continua até que não seja mais vantajoso podar

Abordagens para Medida de Seleção

- Além do ganho de informação, previamente visto, outras medidas podem ser usadas:
 - *Chi-square contingency table statistics (X^2)*
 - Índice de diversidade GINI
 - *Gain-ratio measure*

Dados usados para apresentar as medida de seleção

RADIATION	MENOPAUSE	CLASS
NO	< 60	RECUR
NO	≥ 60	RECUR
NO	< 60	RECUR
NO	NOT	RECUR
YES	≥ 60	NOT RECUR
YES	< 60	NOT RECUR
YES	≥ 60	NOT RECUR
NO	NOT	NOT RECUR
NO	< 60	NOT RECUR
NO	< 60	RECUR

**Valores de dois atributos do banco
de dados de câncer de mama**

		CLASS		
		RECUR	NOT RECUR	
RADIATION	YES	0	3	3
	NO	5	2	7
		5	5	10
		RECUR	NOT RECUR	
AGE OF MENOPAUSE	< 60	3	2	5
	≥ 60	1	2	3
	NOT	1	1	2
		5	5	10

Tabela de Contingência (representação)

VALUE OF ATTRIBUTE	CLASS				TOTAL
	C_1	C_2	\dots	C_c	
A_1	x_{11}	x_{12}		x_{1c}	$x_{1.}$
A_2	x_{21}	x_{22}		x_{2c}	$x_{2.}$
\vdots					\vdots
A_r	x_{r1}	x_{r2}		x_{rc}	$x_{r.}$
	$x_{.1}$	$x_{.2}$	\dots	$x_{.c}$	N

Chi-square contingency table statistics (X^2)

- ❑ Esta é uma medida estatística tradicional para medir a associação entre duas variáveis através da tabela de contingência
- ❑ Ela compara as frequências observadas com as frequências esperadas
- ❑ Quando maior o valor medido, maior a associação.

Chi-square contingency table statistics (χ^2)

- A equação da função é dada por:

$$\chi^2 = \sum \sum \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

- Dado que $E_{ij} = x_{i.}x_{.j}/N$

Os resultados favorecem
o atributo radiação

Para Radiação

$$\chi^2 = \frac{(0 - 1.5)^2}{1.5} + \frac{(3 - 1.5)^2}{1.5} + \frac{(5 - 3.5)^2}{3.5} + \frac{(2 - 3.5)^2}{3.5} = 4.29$$

Para Menopausa

$$\chi^2 = \frac{(3 - 2.5)^2}{2.5} + \dots + \frac{(1 - 1)^2}{1} = 0.533$$

Índice de diversidade GINI

- Bastante similar a medida ganho de informação
- A função GINI mede a *impureza* do atributo em relação a classe
- Dada a probabilidade para cada classe (p_i), a função é dada por:

$$\sum \sum_{j \neq i} p_i p_j = \left(\sum p_i \right)^2 - \sum p_i^2 = 1 - \sum p_i^2$$

Índice de diversidade GINI

- Estima-se a probabilidade de cada classe através de sua frequência relativa ($x_{.i}/N$)
- Assim, a impureza total é:

$$i(t) = 1 - \left(\frac{x_{.1}}{N}\right)^2 - \left(\frac{x_{.2}}{N}\right)^2 - \dots$$

- E a impureza da linha A_1 é:

$$i(A_1) = 1 - \left(\frac{x_{11}}{x_{1.}}\right)^2 - \left(\frac{x_{12}}{x_{1.}}\right)^2 - \dots$$

Índice de diversidade GINI

- O aumento na impureza é dado por:

Impureza da classe
menos (-)
Média ponderada da impureza das linhas

$$\begin{aligned}i &= i(t) - \frac{x_{1.}}{N} i(A_1) - \frac{x_{2.}}{N} i(A_2) - \dots \\&= \left(1 - \sum \left(\frac{x_{.j}}{N}\right)^2\right) - \frac{x_{1.}}{N} \left(1 - \sum \left(\frac{x_{1j}}{x_{1.}}\right)^2\right) - \dots \\&= \frac{1}{N} \left(\sum \sum \frac{x_{ij}^2}{x_{i.}} - \sum \frac{x_{.j}^2}{N}\right).\end{aligned}$$

Índice de diversidade GINI

Para Radiação

$$i = \frac{1}{10} \left(\frac{0^2}{3} + \frac{3^2}{3} + \frac{5^2}{7} + \frac{2^2}{7} \right) - \left(\frac{5^2}{10} + \frac{5^2}{10} \right) = 0.21429.$$

Para Menopausa

$$i = \frac{1}{10} \left(\frac{3^2}{5} + \frac{2^2}{5} + \dots \right) - \left(\frac{5^2}{10} + \frac{5^2}{10} \right) = 0.026667$$

Gain-ratio measure

- Essa medida incorpora a noção de que o atributo possui informação

$$GR(A) = \frac{IM(A)}{IV(A)}$$

- O $IV(A)$ é a informação do atributo A

$$IV(A) = - \sum \frac{x_{i.}}{N} \log \left(\frac{x_{i.}}{N} \right)$$

- Essa função tem valores altos quando os exemplos estão dispersos e baixo caso contrário

Gain-ratio measure

Para Radiação

$$IV = -\frac{3}{10} \log \frac{3}{10} - \frac{7}{10} \log \frac{7}{10} = 0.61086$$

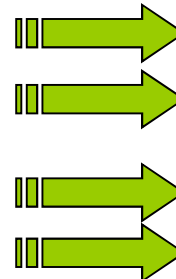
$$GR = \frac{0.27436}{0.61086} = 0.44913$$

Para Menopausa

$$IV = -\frac{5}{10} \log \frac{5}{10} - \frac{3}{10} \log \frac{3}{10} - \frac{2}{10} \log \frac{2}{10} = 1.02965$$

$$GR = \frac{0.02740}{1.02965} = 0.02662$$

Comparação entre as medidas de seleção



MEASURE	RADIATION	MENOPAUSE	RATIO
IM	0.27436	0.02740	10.01
χ^2	4.29	0.533	8.05
G	5.49	0.548	10.01
GINI	0.21429	0.02667	8.03
GAIN RATIO	0.44913	0.02661	16.88
MARSH	0.23046	0.02219	10.38

Referências

- Tom Mitchell. *Machine Learning*. McGraw-Hill. 1997.
- John Mingers. *An Empirical Comparison of Pruning Methods for Decision Tree Induction*. **Machine Learning**, 4, 227-243, 1989.
- John Mingers. *An Empirical Comparison of Selection Measures for Decision-Tree Induction*. **Machine Learning**, 3: 319-342, 1989.