

Fundamentos de Inteligência Artificial [5COP099]



Dr. Sylvio Barbon Junior
Saulo Martiello Mastelini

Departamento de Computação - UEL

1º Semestre

Assunto

Aula 11

Modelos Preditivos - Árvore de Decisão (corte)

Sumário

- Poda em Árvores de Decisão
- Métodos de Proda
- Pré-poda
- Pos-poda

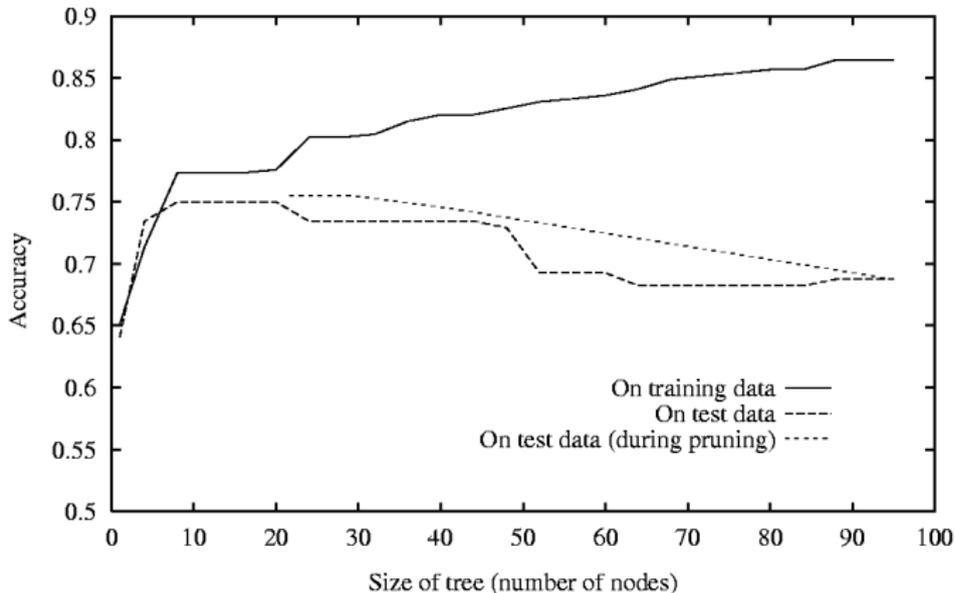
Poda em Árvores de Decisão

- A poda é uma parte das mais importantes do processo de construção da árvore, principalmente em domínios ruidosos.
- Os ruídos causam dois problemas:
 - Podem classificar novos objetos em um modo não confiável;
 - Uma árvore induzida tende a ser grande e difícil de interpretação;
- Os nós mais profundos tem baixa importância pois poucas amostras são classificadas pelas folhas destes;
- Nós mais profundos refletem mais o conjunto de treinamento (overfitting - superajustamento) e aumentam o nível erro devido a variância.

Poda em Árvores de Decisão

- Podar uma árvore é trocar nós mais profundos por folhas, minimizando os problemas.
- Podar uma árvore de decisão certamente irá causar a classificação incorreta de alguns exemplos de treinamento, porém é eficiente na avaliação de novos exemplos.

Poda em Árvores de Decisão



Métodos de Poda

- Podar uma árvore é trocar nós mais profundos por folhas, minimizando os problemas.
- Podar uma árvore de decisão certamente irá causar a classificação incorreta de alguns exemplos de treinamento, porém é eficiente na avaliação de novos exemplos.
- Os métodos de poda pode ser divididos em dois grupos:
 - **Pré-poda:** durante a construção da árvore quando algum critério é atingido;
 - **Pós-poda:** após a construção observado certos critérios.
- O importante é estimar o ponto de equilíbrio entre a árvore e a estimativa de erro.

Métodos de Poda - Pré-poda

- Este tipo de poda conta com regras de para que previnem a construção daqueles ramos que não melhoram a precisão preditiva da árvore;
- Tem a vantagem de economia de tempo da montagem da árvore;
- Normalmente temos as seguintes regras de parada:
 - Todas as observações de um nó pertencem a uma mesma classe;
 - Todas as observações alcançando um nó tem mesmo vetor de características (mas não necessariamente pertencem a mesma classe)
 - O número de observações é menor que um certo limiar;
 - O mérito atribuído a todos os possíveis testes que particionaram o conjunto de observações é muito baixo;

Métodos de Poda - Pós-poda

- Este é o método mais comum para as Árvores de Decisão;
- Segundo Quinlan (1988) "Construir e podar uma árvore é lento, mas mais confiável";
- Um dos métodos mais simples é baseado em **duas medidas**:
 - Erro estático: É o número de classificações incorretas considerando que todos os exemplos que chegam nesse nó são classificados usando a classificação majoritária da distribuição de classes deste nó;
 - Erro de Backed-up: É a soma das classificações incorretas de todas as subárvores do nó corrente;
- Se o erro de backed-up é maior ou igual ao erro estático, então o nó é trocado pela folha com a classe majoritária;

Métodos de Poda - Pós-poda

- A poda **por custo de complexidade** é um dos mais utilizados;
- Inicialmente é gerada uma árvore completa e conseqüentemente uma sequência de árvores cada vez menores, sendo escolhida uma das subárvores;
- Este método é baseado em dois parâmetros: taxa de erro $R(T)$ e tamanho da árvore $|T|$ em termos de folhas.
- A medida de custo complexidade é: $R_\alpha(T) = R(T) + \alpha|T|$;
- O valor de α é a importância relativa do tamanho da árvore em relação à taxa de erro;
- A árvore selecionada é a que apresentar o menor $R_\alpha(T)$;
- Este é o core de funcionamento da árvore **CART**.

Referências

1. Coppin, B. Inteligência Artificial. LTC. 2010.
2. Russell, S.; Norvig, P. Artificial Intelligence: a modern approach. Prentice Hall. 2010. Localização: BC – Número de Chamada: 519.683 R967a 3.ed.
3. Luger, G. F. Inteligência Artificial: estruturas e estratégias para a resolução de problemas complexos. Bookman. 2004. Localização: BC – Número de Chamada: 519.683 L951a 4.ed.
4. Carvalho, André, et al. "Inteligência Artificial—uma abordagem de aprendizado de máquina." Rio de Janeiro: LTC (2011).