Lista 03 - IA

Luiza Ávila

01- O artigo “A comparative study of decision tree ID3 and C4.5” tem como objetivo principal colocar os algoritmos de árvore de decisão ID3 e C4.5 frente a frente e compará-los. Primeiro, somo introduzidos ao algoritmo ID3, sua funcionalidade é explicada e ele é exemplificado para dizer se “devem jogar bola” ou não. Logo após, somos introduzidos ao algoritmo de C4.5, o mesmo exemplo é usado nele e acrescenta-se um atributo numérico no exemplo, para demonstrar sua execução. Depois os autores comparam os dois algoritmos usando dados coletados pelo exemplo, o C4.5 é comparado com o C5.0 e o C5.0 com o CART (só na teoria, nenhuma exemplificação foi mostrada), para, por fim, os autores concluírem que o C4.5 é o melhor algoritmo a ser usado.

02-

A) Método de amostragem que divide a base em X folds e usará X-1 deles para treinar e validará com um deles. Ele fará esse processo X vezes, cada vez usando 1 deles para validar. A média então é calculada.

B) Caso especial de Cross Validation.

C) Método de amostragem que serve para separar X% para treinamento e (100-X)% para teste. Para pequenos arquivos, nem sempre é possível separar uma parte dos exemplos. Para uma amostra de tamanho n uma hipótese é induzida utilizando (n-1) exemplos; a hipótese é então testada no único exemplo remanescente. Este processo é repetido n vezes, cada vez induzindo uma hipótese deixando de considerar um único exemplo. O erro é a soma dos erros em cada teste dividido por n.

03-

A) Flexibilidade, interpretabilidade

B) Valores ausentes, instabilidade

04-

A) Particionando o conjunto de treinamento nos valores deste atributo levará a um grande número de subconjuntos, cada um contendo somente um caso.Como todos os subconjuntos (de 1 elemento) necessariamente contêm exemplos de uma mesma classe, info(ID,T)=0, assim o ganho de informação deste atributo será máximo. Para solucionar esse problema, usamos a razão de ganho. Então, a razão de ganho expressa a proporção de informação gerada pela partição que é útil, ou seja, que aparenta ser útil para a classificação. A razão de ganho é = ganho/entropia.

B) Só lida com atribtuos discretos; nenhuma forma de tratar valores deconhecidos; não tem algoritmo pós-poda.

C) O C4.5 é uma extensão do algoritmo ID3. Ele lida com algoritmos contínuos e discretos, lidar com atributos incompletos, poda de árvores e usa razão de ganho.

D)

1. Começar com todos os exemplos de treino;
2. Escolher o teste (atributo) que melhor divide os exemplos, ou seja agrupar exemplos da mesma classe ou exemplos semelhantes;
3. Para o atributo escolhido, criar um nó filho para cada valor possível do atributo;
4. Transportar os exemplos para cada filho tendo em conta o valor do filho;
5. Repetir o procedimento para cada filho não "puro". Um filho é puro quando cada atributo X tem o mesmo valor em todos os exemplos.