



Unyleya
EDUCACIONAL



ID3

O algoritmo ID3 foi proposto por Quilan, o mesmo que apresentou o conceito das árvores de decisão.

O seu funcionamento é baseado nos sistemas de inferência e nos conhecimentos acerca do processo de aprendizagem.

A partir de um conjunto de elementos esse algoritmo extrai informações e as internaliza para resolução de problemas futuros.

Utilizando o ganho de informação, o ID3 identifica o atributo mais importante dentro do conjunto de dados oferecido para treinamento.

O algoritmo utiliza o atributo escolhido para dividir o conjunto de treinamento em partes menores e específicas para as diferentes classes.

O atributo mais importante é posicionado no topo do árvore. A partir de então a árvore é construída de cima para baixo.

A recursão é utilizada para sempre criar o próximo nó com o atributo que representa o maior ganho de informação.

Os passos do ID3

1. Inicializa o conjunto de dados para treinamento;
2. Escolhe o atributo de maior ganho de informação e agrupa elementos por classes;
3. Com o atributo escolhido se cria um nó folha para cada opção de valor do atributo;
4. Transporta exemplos para o filho, de acordo com sua característica;
5. Repete o procedimento para todos os filhos que não são “puros”.

Um filho é puro sempre que cada atributo X tem valor igual para todos os exemplos.

O funcionamento do algoritmo ID3 não garante que a solução ótima seja encontrada. Isso significa que ele pode encontrar soluções com ótimos locais em que a árvore não tem a melhor estrutura possível.

O seu conceito de selecionar sempre para próxima opção o melhor atributo possível indica que ele trabalha com um conceito semelhante ao dos algoritmos gulosos.

Pode ser adaptado para melhorar o seu processo de construção das árvores.

Utilizando o ID3 pode acontecer o overfitting, isso significa que o algoritmo não tem um controle sobre o tamanho das árvores que gera e por isso pode se especializar demais nos dados de treinamento.

Não há garantias de que o ID3 gerará a menor árvore de decisão possível.

Quando utilizado com dados contínuos, o desempenho do algoritmo ID3 tende a piorar.

Isso acontece porque o algoritmo terá mais dificuldades para encontrar o ponto de divisão dos atributos que possuírem muitos valores contínuos.

O ganho da informação no algoritmo ID3 é calculado utilizando a fórmula da entropia.



Obrigada!

Ana Laurentino