Mineração de Dados

Trabalho 5

Arthur do Prado Labaki - 11821BCC017

21-04, 2023

GBC212

Estudo do Conceito

Uma árvore de decisão é um modelo de mineração de dados que usa uma estrutura em forma de árvore para representar uma série de decisões e suas possíveis consequências. Cada nó da árvore representa uma escolha, com as ramificações indicando as possíveis consequências dessa escolha. As árvores de decisão são usadas em uma variedade de aplicações, incluindo classificação e previsão, e são especialmente úteis quando se trata de problemas complexos e dados com muitas variáveis.

As árvores de decisão começam com um nó raiz que representa o conjunto completo de dados. Cada nó subsequente representa uma variável e uma decisão baseada nessa variável. Essas decisões são representadas por ramos da árvore que levam a novos nós. Os nós finais da árvore são chamados de nós folha e representam as classes ou valores previstos para uma determinada instância de dados.

As árvores de decisão podem ser construídas por meio de vários algoritmos, incluindo o ID3 (Iterative Dichotomiser 3), o C4.5 e o CART (Classification and Regression Trees). Esses algoritmos usam técnicas de divisão recursiva para escolher a melhor variável para dividir os dados em cada nó da árvore, de modo a maximizar a pureza dos subconjuntos de dados resultantes. O processo de construção da árvore continua até que um critério de parada seja atingido, como um número mínimo de instâncias por nó folha ou uma profundidade máxima da árvore.

Resolução do Exercício 1)

Dado a base de dados abaixo, encontre a arvore de decisão resultante da maximização do ganho de informação (Algoritmo ID3)

Dia	Panorama	Temperatura	Umidade	Vento	Jogar Tênis
1	Ensolarado	Quente	Alta	Fraco	Não
2	Ensolarado	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuvoso	Intermediária	Alta	Fraco	Sim
5	Chuvoso	Fria	Normal	Fraco	Sim
6	Chuvoso	Fria	Normal	Forte	Não
7	Nublado	Fria	Normal	Forte	Sim
8	Ensolarado	Intermediária	Alta	Fraco	Não
9	Ensolarado	Fria	Normal	Fraco	Sim
10	Chuvoso	Intermediária	Normal	Fraco	Sim
11	Ensolarado	Intermediária	Normal	Forte	Sim
12	Nublado	Intermediária	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chuvoso	Intermediária	Alta	Forte	Não

$$E(T) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad E(T,X) = \sum_{c \in X} P(c)E(c) \qquad \textbf{Gain(T,X) = E(T) - E(T,X)}$$

Figura 1: Base de dados

Primeiramente precisamos calcular o ganho de informação de cada atributo. Vamos começar com o atributo Panorama. Para calcular o ganho de informação, precisamos primeiro calcular a entropia do conjunto de dados original (Formula E(T) na imagem):

$$E(S) = -(9/14) * log2(9/14) - (5/14) * log2(5/14) = 0.940$$

Agora, para o atributo Panorama, temos três valores possíveis: Ensolarado, Nublado e Chuvoso. Vamos calcular a entropia de cada subconjunto resultante da divisão pelo atributo Panorama e, em seguida, calcular a entropia ponderada média (Formula E(T,X) na imagem):

$$E(Ensolarado) = -(2/5) * log2(2/5) - (3/5) * log2(3/5) = 0.971$$

$$E(Nublado) = -(4/4) * log2(4/4) - (0/4) * log2(0/4) = 0$$

$$E(Chuvoso) = -(3/5) * log2(3/5) - (2/5) * log2(2/5) = 0.971$$

$$E(Panorama) = (5/14)*E(Ensolarado)+(4/14)*E(Nublado)+(5/14)*E(Chuvoso) = 0.693$$

O ganho de informação para o atributo Panorama é, portanto (Formula Gain(T,X) na imagem):

$$Ganho(Panorama) = E(S) - E(Panorama) = 0.940 - 0.694 = 0.246$$

Agora, basta repetir as contas para todos os outros atributos (Temperatura, Umidade e Vento). Com isso temos:

Temperatura:

$$E(Quente) = -(2/4) * log2(2/4) - (2/4) * log2(2/4) = 1.000$$

$$E(Intermediária) = -(4/6) * log2(4/6) - (2/6) * log2(2/6) = 0.918$$

$$E(Fria) = -(3/4) * log2(3/4) - (1/4) * log2(1/4) = 0.811$$

$$E(Temperatura) = (4/14) * E(Quente) + (6/14) * E(Moderada) + (4/14) * E(Fria) = 0.911$$

$$Ganho(Temperatura) = E(S) - E(Temperatura) = 0.029$$

Umidade:

$$E(Alta) = -(3/7) * log2(3/7) - (4/7) * log2(4/7) = 0.985$$

$$E(Normal) = -(6/7) * log2(6/7) - (1/7) * log2(1/7) = 0.592$$

$$E(Umidade) = (7/14) * E(Alta) + (7/14) * E(Normal) = 0.789$$

$$Ganho(Umidade) = E(S) - E(Umidade) = 0.151$$

Vento:

$$E(Fraco) = -(6/8) * log2(6/8) - (2/8) * log2(2/8) = 0.811$$

$$E(Forte) = -(3/6) * log2(3/6) - (3/6) * log2(3/6) = 1.000$$

$$E(Vento) = (8/14) * E(Fraco) + (6/14) * E(Forte) = 0.892$$

$$Ganho(Vento) = E(S) - E(Vento) = 0.048$$

Com base nos cálculos de ganho de informação para cada atributo, podemos construir a árvore de decisão:

Começando pelo nó raiz, o atributo que apresentou o maior ganho de informação foi o "Tempo", com valor de ganho igual a 0.246. Dessa forma, o primeiro nó da árvore será dividido em três ramos, correspondendo às possibilidades de valores para esse atributo: Ensolarado, Nublado e Chuvoso.

Em seguida, cada um dos ramos será avaliado de forma separada, utilizando o mesmo critério de seleção do atributo com maior ganho de informação. No ramo correspondente ao valor Ensolarado do atributo Tempo, será feito novamente o calculo da informação, sendo :

$$Ganho(Umidade) = E(S) - E(Umidade) = 0.971 - 0.940 = 0.031$$
$$Ganho(Temperatura) = E(S) - E(Temperatura) = 0.971 - 0.693 = 0.278$$
$$Ganho(Vento) = E(S) - E(Vento) = 0.971 - 0.892 = 0.079$$

Como o atributo Umidade apresenta o maior ganho de informação (0.031), seguido pelo atributo Vento (0.079) e pelo atributo Temperatura (0.278). Portanto, o atributo Umidade é escolhido como o atributo de divisão para o nó correspondente ao valor Ensolarado do atributo Tempo. E esse ramo é dividido em dois ramos, correspondentes aos valores de Umidade Alta e Normal.

Para o ramo correspondente ao valor Nublado do atributo Tempo, não há necessidade de selecionar um atributo adicional, já que todos os exemplos pertencem à mesma classe (Jogar), portanto esse nó é folha, indicando que, independentemente das outras condições, quando o tempo está nublado, o jogo sempre ocorre.

Finalmente, para o ramo correspondente ao valor Chuvoso do atributo Tempo, refazendo as contas, temos:

$$Ganho(Umidade) = E(S) - E(Umidade) = 0 - \textit{Já escolhido}$$
$$Ganho(Temperatura) = E(S) - E(Temperatura) = 0.971 - 0.911 = 0.060$$
$$Ganho(Vento) = E(S) - E(Vento) = 0.971 - 0.892 = 0.079$$

Então o atributo que apresentou o maior ganho de informação foi o atributo Vento, com valor de ganho igual a 0.079. Dessa forma, esse ramo é dividido em dois ramos, correspondentes aos valores de Vento Fraco e Forte.

O atributo Temperatura não foi selecionado em nenhum momento, pois apresentou o menor ganho de informação entre todos os atributos, indicando que ele não é um bom indicador para separar as classes de Jogar ou Não Jogar.

Por fim, temos a arvore:

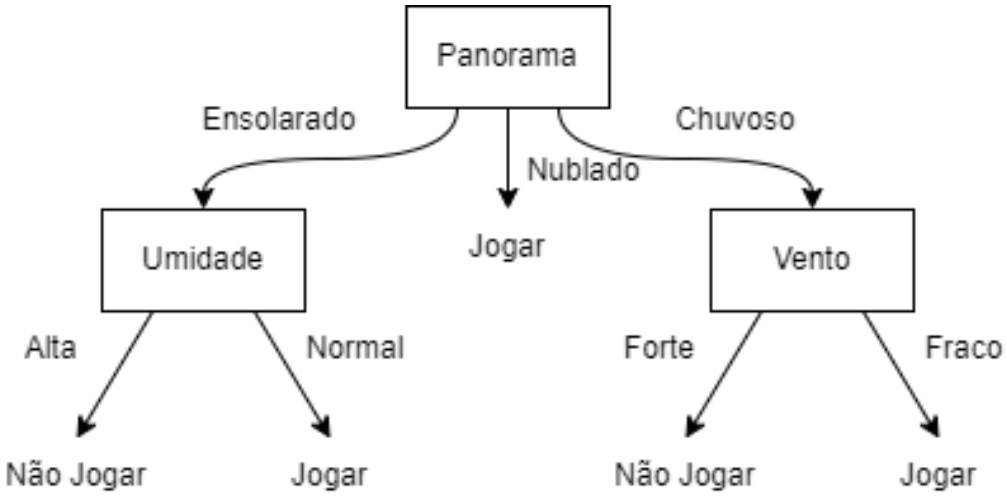

Figura 2: DÁrvore de decisão