Universidade Estadual do Oeste do Paraná
Inteligência Artificial
Docentes: Huei D. Lee e Newton Spolaor
Isabela Pimentel Loebel

TRABALHO 3: GANHO MÁXIMO, RAZÃO DE GANHO E ÍNDICE DE GINI

1. Introdução

O objetivo do presente trabalho é apresentar e conceituar os temas propostos, sendo eles: ganho máximo, razão de ganho e índice de Gini, assim como ressaltar sua significância no contexto a ser aplicado. Tais temas estão relacionados ao cálculo da importância de atributos presentes em uma base de dados.

2. Ganho Máximo

Como define Baranauskas (s.d.), o ganho máximo seleciona o atributo que possui o maior ganho de informação esperado, isto é, seleciona o atributo que resultará no menor tamanho esperado das subárvores, assumindo que a raiz é o nó atual.

Partindo do conceito disposto, o ganho máximo é um conceito utilizado em algoritmos de árvore de decisão para determinar o melhor atributo para dividir os dados em cada nó. O atributo que maximiza o ganho de informação é escolhido, o qual é uma medida da redução da entropia ou desordem nos dados após a divisão.

Quinlan (1986) fornece uma explicação detalhada do algoritmo ID3 e seu uso do ganho máximo em sua obra, mas sucintamente, afirma-se que a determinação do atributo ótimo para particionar o conjunto de exemplos em um algoritmo de árvore de decisão é realizada por meio do cálculo do ganho de informação associado a cada atributo. Este ganho é obtido pela subtração da entropia de todo o conjunto pela entropia de cada atributo, podendo ser expresso pela equação 1.

$$Ganho\ de\ Informação\ (A, D) = H(D) - H(A, D)$$

Equação 1 – Ganho da informação.

Onde:

- $H(D)$ é a entropia em relação ao total de exemplos;
- $H(A, D)$ é a entropia em relação ao atributo A

Essa abordagem avalia a redução na incerteza do conjunto de dados que pode ser alcançada ao considerar diferentes atributos para a divisão dos dados. O atributo que, entre todos os utilizados na classificação das amostras, proporcionar o maior ganho de informação será o atributo com ganho máximo, ou seja, o mais relevante.

3. Razão de Ganho

A Razão de Ganho (*Gain Ratio*), conforme definida por Baranauskas (s.d), representa a proporção de informação resultante da partição que é considerada útil para a classificação. Este conceito pode ser entendido como uma variação do Ganho de Informação que leva em conta não apenas a redução da entropia, mas também o número de categorias ou classes presentes no atributo, que nada mais é do que o ganho de informação relativo. A utilização da Razão de Ganho é proposta por Quinlan (1986) no contexto do algoritmo C4.5, visando mitigar o viés em direção a atributos com muitas categorias.

O Ganho de Informação é um critério comumente empregado em algoritmos de árvore de decisão para determinar qual atributo proporciona a melhor divisão dos dados. No entanto, em situações em que os atributos possuem um número desigual de categorias, o Ganho de Informação pode favorecer atributos com mais categorias, levando a uma potencial distorção nos resultados do modelo. Para contornar esse problema, a Razão de Ganho é calculada, considerando não apenas a redução da entropia, mas também o número de categorias presentes em cada atributo. Dessa forma, a Razão de Ganho busca equilibrar a influência dos diferentes atributos na construção da árvore de decisão, garantindo uma abordagem mais justa e eficaz na seleção dos atributos para particionar o conjunto de dados.

A expressão utilizada para calcular a razão de ganho é uma modificação da equação do Ganho de Informação, adaptada para considerar o número de categorias ou classes presentes em cada atributo, podendo ser expressa pela equação 2.

$$Razão\ de\ Ganho(A) = \frac{Ganho\ da\ Informação(A)}{Informação\ Intríseca(A)}$$

Equação 2 – Razão de ganho.

Onde:

- A é o atributo em consideração;
- $Ganho\ da\ Informação(A)$ é o ganho de informação associado ao atributo A, calculado conforme a equação 1;

- $Informação\ Intríseca(A)$ representa a informação intrínseca do atributo A, que é calculada como a entropia média dos valores possíveis de A, ponderada pela frequência de ocorrência de cada valor.

Essa relação entre o ganho de informação e a informação intrínseca permite avaliar não apenas a redução da entropia proporcionada pelo atributo, mas também a sua complexidade, de modo a evitar o viés em direção a atributos com muitas categorias, expressando a proporção de informação gerada pela partição que é útil, ou seja, que aparenta ser mais importante para a classificação.

4. Índice de Gini

O Índice de Gini, proposto em 1992 por Corrado Gini, estatístico italiano, é uma medida adicional de impureza ou desordem nos dados, comumente utilizada na análise de árvores de decisão. Além disso, o índice de Gini é empregado em análises econômicas e sociais para avaliar a distribuição de renda em uma determinada população ou país.

Este índice é um indicador de dispersão estatística que quantifica a heterogeneidade dos dados. Sua aplicação na seleção de atributos reside na escolha do melhor atributo para particionar os dados em um algoritmo de árvore de decisão. O atributo que minimiza o índice de Gini é selecionado como critério de divisão para a árvore, podendo, o índice, ser expresso pela equação 3.

$$Gini(S) = i - \sum_{i=1}^{k} p(c_i\ |n)^2$$

Equação 3 – Índice de Gini.

Onde:

- S é o conjunto de dados;
- n é o número de registros no conjunto S;
- c_i é a classe relativa ao n;
- p_i é a probabilidade relativa da classe em c_i em S;
- k é o número de classes.

Essa equação indica se há homogeneidade dos dados (todos os exemplos pertencem à mesma classe) ou heterogeneidade (os exemplos estão igualmente distribuídos entre todas as classes).

5. Conclusão

Os tópicos propostos demonstraram ser de suma importância para a seleção de atributos e a subsequente construção de um modelo analítico. Esses tópicos possibilitam a quantificação numérica do impacto de um atributo específico em relação ao conjunto total de dados. Isso, por sua vez, influencia diretamente a eficácia do modelo construído.

A seleção de atributos é um passo crítico no processo de modelagem, pois permite que os analistas identifiquem e compreendam as variáveis que têm o maior impacto na previsão ou classificação de um resultado. Ao quantificar o impacto de um atributo, podemos avaliar sua relevância e utilidade para o modelo, o que pode levar a melhorias significativas na precisão e eficiência do modelo.

Portanto, a importância dos temas propostos não pode ser subestimada, pois eles desempenham um papel fundamental na otimização do processo de modelagem e na maximização do valor extraído do conjunto de dados.

6. Referências Bibliográficas

BARANAUSKAS, J. **Indução de Árvores de Indução de Árvores de Decisão.** [s.d.]. FFCLRP-USP. Disponível em: <https://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-Arvores-Decisao.pdf>. Acesso em: 23 de abr de 2024.

QUINLAN, J. R. Decision Trees: Induction Algorithms and Applications. 1986. Vol 1. Disponível em: < https://link.springer.com/article/10.1007/BF00116251>. Acesso em: 23 de abr de 2024.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 2011. 3a ed. Disponível em: < https://books.google.com.br/books?id=pQws07tdpjoC&printsec=frontcover&hl=pt-BR#v=onepage&q&f=false>. Acesso em: 23 de abr de 2024.

HOEHSTEIN, G. Algoritmo ID3 e C.45. Disponível em: <https://pt.slideshare.net/iaudesc/algoritmoid3ec45gilcimar>. Acesso em: 23 de abr de 2024.