Nguyễn Thị Duyên – 20211423 – DCCNTT12.10.6

**I. Thuật toán Cart (Classification and Regression Trees)**

**1. Thuật toán Cart là gì**

Thuật toán CART (cây phân loại và hồi quy - Classification and Regression Tree), là một thuật toán được Leo Breiman giới thiệu năm 1984, dùng để xây dựng cây quyết định đa năng, với mục đích phân loại và hồi quy dữ liệu.

Thuật toán CART là nền tảng cho những thuật toán khác mạnh mẽ hơn như bagged decision trees (tập hợp nhiều cây quyết định để cùng đoán nhận), random forest (rừng ngẫu nhiên) và boosted decision trees (tạp hợp các cây khác nhau, là các model; và một siêu model - meta model - để tập hợp kết quả đoán nhận của các model con).

**2. Tiêu chí phân chia**

Gini Index chính là chỉ số đo lường mức độ đồng nhất hay nhiễu loạn của thông tin, hay sự khác biệt về các giá trị mà mỗi điểm dữ liệu trong 1 tập con, hoặc một nhánh của DT. Công thức của Gini index có thể dùng cho cả dữ liệu rời rạc và liên tục. Nếu điểm dữ liệu thuộc về 1 nút và có chung thuộc tính bất kỳ thì nút này thể hiện sự đồng nhất lúc này Gini = 0 ngược lại Gini sẽ lớn.

Công thức tổng quát của Gini index



Trong đó pi là tỉ lệ của các lớp i trong tập hợp. Gia trị Gini nằm trong khoảng từ 0 đến 1.

## **3. Phương pháp phân chia**

CART thực hiện chia nhị phân tại mỗi nút, có nghĩa là nó chỉ tạo ra hai nhánh con cho mỗi quyết định. Do đó, dù tập dữ liệu gốc có bao nhiêu lớp (class) thì mỗi lần phân chia sẽ chỉ tạo ra hai tập con.

## **4. Quy trình xây dựng**

- Tại mỗi nút, CART duyệt qua tất cả các đặc trưng và các ngưỡng cắt để tìm ra điểm phân chia tối ưu dựa trên Chỉ số Gini hoặc Sai số Bình phương Trung bình.

- Sau đó, cây tiếp tục được xây dựng bằng cách chia đôi dữ liệu dựa trên đặc trưng và ngưỡng cắt được chọn cho đến khi đạt đến độ sâu tối đa hoặc một tiêu chí dừng khác.

**5. Ưu điểm và nhược điểm của thuật toán Cart**

|  |  |
| --- | --- |
| **Ưu điểm** | **Nhược điểm** |
| Dễ hiểu và trực quan (Cây quyết định có thể được trực quan hóa, giúp người dùng dễ dàng hiểu cách mà thuật toán đưa ra quyết định). | Nhạy cảm với dữ liệu nhiễu (CART có thể tạo ra các cây phức tạp nếu dữ liệu chứa nhiều nhiễu, dẫn đến overfitting). |
| Không yêu cầu tiền xử lý phức tạp (CART không yêu cầu chuẩn hóa dữ liệu hoặc xử lý các giá trị thiếu, vì nó có thể làm việc với dữ liệu thô). | Cây quyết định có thể quá lớn (Nếu không có cắt tỉa hợp lý, cây quyết định có thể trở nên quá lớn và khó hiểu). |
| Có khả năng xử lý dữ liệu phân loại và hồi quy (CART có thể được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression)). | Không đảm bảo tìm kiếm toàn cục (Thuật toán không đảm bảo rằng nó tìm ra cây tối ưu, mà chỉ tìm kiếm trong không gian khả thi). |
| Tự động chọn thuộc tính quan trọng (Thuật toán tự động xác định các thuộc tính quan trọng để chia tách, giúp giảm thiểu sự can thiệp của người dùng). | Thiếu tính chính xác trong một số trường hợp (CART có thể không chính xác trong các bài toán phức tạp hơn, đặc biệt là khi có nhiều biến tương tác). |
| Khả năng xử lý dữ liệu phi tuyến (CART có thể mô hình hóa các mối quan hệ phi tuyến giữa các biến). | Khó khăn trong việc xử lý dữ liệu không cân bằng (Nếu có sự phân bố không đồng đều giữa các lớp trong dữ liệu, CART có thể thiên lệch về lớp có số lượng lớn hơn). |
| Khả năng cắt tỉa (Cắt tỉa cây giúp giảm thiểu overfitting và cải thiện khả năng tổng quát của mô hình). | Khó khăn trong việc dự đoán liên tục (Mặc dù CART có thể thực hiện hồi quy, nhưng nó có thể không chính xác như các phương pháp hồi quy khác như hồi quy tuyến tính hoặc hồi quy Ridge). |

**II. ID3 (Iterative Dichotomiser 3)**

1. **Khái niệm**

ID3 là một thuật toán cây quyết định được phát triển bởi Ross Quinlan, chủ yếu được sử dụng cho bài toán phân loại. ID3 sử dụng khái niệm Độ lợi Thông tin (Information Gain), một thước đo dựa trên Entropy, để lựa chọn cách chia dữ liệu tại mỗi nút của cây

**2. Tiêu chí phân chia**

- ID3 chọn đặc trưng tại mỗi nút dựa trên Độ lợi Thông tin. Độ lợi Thông tin xác định sự giảm độ bất định trong phân phối lớp khi một đặc trưng được sử dụng để phân chia dữ liệu.

- Entropy là một thước đo mức độ hỗn loạn hoặc không chắc chắn trong dữ liệu. Giá trị entropy cao biểu thị tập dữ liệu hỗn loạn, không thuần nhất, và ngược lại, entropy bằng 0 biểu thị dữ liệu thuần nhất (tất cả điểm dữ liệu thuộc cùng một lớp).

**3. Phương pháp phân chia**

ID3 có thể thực hiện phân chia đa nhánh (multisplits) tại mỗi nút, tức là số nhánh có thể lớn hơn hai nếu đặc trưng được sử dụng để phân chia có nhiều giá trị.

**4. Quy trình xây dựng**

- ID3 tính entropy cho tập dữ liệu hiện tại

- Sau đó, nó tính Độ lợi Thông tin cho mỗi đặc trưng và chọn đặc trưng có Độ lợi Thông tin lớn nhất để phân chia tại nút.

- Quy trình này tiếp tục lặp lại cho đến khi cây đạt đến một tiêu chí dừng, như tất cả các điểm dữ liệu thuộc một lớp, hoặc không còn đặc trưng nào để phân chia

**5. Ưu điểm và nhược điểm của ID3**

|  |  |
| --- | --- |
| **Ưu điểm** | **Nhược điểm** |
| Dễ hiểu và trực quan (Cây quyết định do ID3 tạo ra rất dễ hiểu và trực quan, giúp người dùng dễ dàng theo dõi quy trình ra quyết định). | Nhạy cảm với dữ liệu nhiễu (ID3 có thể tạo ra cây phức tạp khi dữ liệu có nhiễu, dẫn đến hiện tượng overfitting). |
| Tốc độ nhanh (ID3 có thể xây dựng cây quyết định nhanh chóng, nhất là trên các bộ dữ liệu nhỏ đến trung bình). | Quá lớn và khó hiểu (Cây quyết định có thể trở nên rất lớn nếu không có cắt tỉa hợp lý, gây khó khăn cho việc diễn giải và sử dụng). |
| Không yêu cầu tiền xử lý phức tạp (ID3 không yêu cầu chuẩn hóa dữ liệu và có thể làm việc với các giá trị thiếu). | Khó khăn trong việc xử lý thuộc tính liên tục (Mặc dù ID3 có thể xử lý các thuộc tính rời rạc, nhưng việc xử lý các thuộc tính liên tục đòi hỏi thêm bước chuyển đổi thành dạng phân loại). |
| Tự động chọn thuộc tính tốt nhất(ID3 sử dụng chỉ số thông tin (Information Gain) để chọn thuộc tính phân chia tốt nhất, giúp mô hình tập trung vào các yếu tố quan trọng). | Thiếu tính chính xác với các lớp không cân bằng (ID3 có thể thiên lệch về lớp có số lượng lớn hơn, dẫn đến kết quả không chính xác trong các bài toán phân loại không cân bằng). |
| Khả năng xử lý các thuộc tính phân loại (ID3 có thể xử lý tốt các thuộc tính phân loại và không yêu cầu chuyển đổi dữ liệu thành dạng số). | Cắt tỉa cây không tự động (ID3 không có cơ chế cắt tỉa tự động, do đó, cần phải thực hiện cắt tỉa bằng tay hoặc sử dụng phương pháp khác để cải thiện độ chính xác). |