**Trương Khánh Linh – 20211553**

**DCCNTT12.10.6**

**1. Quy trình phân lớp ảnh với cây quyết định**

Phân lớp ảnh với cây quyết định là quá trình xây dựng mô hình phân lớp dựa trên việc học các quy tắc ra quyết định từ dữ liệu huấn luyện. Trong quy trình này, các ảnh thường được chuyển đổi thành các dạng dữ liệu số để cây quyết định có thể phân tích và đưa ra quyết định.

**Bước 1: Tiền xử lý và trích xuất đặc trưng ảnh**

* **Tiền xử lý**: Trước khi đưa vào cây quyết định, ảnh thường cần được chuyển đổi hoặc tiền xử lý như thay đổi kích thước, lọc nhiễu, chuẩn hóa.
* **Trích xuất đặc trưng**: Ảnh được chuyển thành các vector đặc trưng, chứa các thông tin quan trọng giúp mô hình dễ phân biệt. Một số kỹ thuật trích xuất đặc trưng phổ biến là:
  + **HOG (Histogram of Oriented Gradients)**: Trích xuất thông tin về gradient của ảnh.
  + **SIFT (Scale-Invariant Feature Transform)**: Tạo ra các đặc trưng bất biến theo tỷ lệ và xoay ảnh.
  + **Các mô hình mạng nơ-ron tích chập (CNN)**: Trích xuất đặc trưng ảnh sâu qua các lớp tích chập.

**Bước 2: Chia bộ dữ liệu**

Dữ liệu cần được chia thành 2 tập:

* **Tập huấn luyện**: Tập dùng để xây dựng mô hình.
* **Tập kiểm thử**: Tập dùng để đánh giá hiệu suất của mô hình.

**2. Phương pháp phân lớp với cây quyết định**

Cây quyết định là một thuật toán phân lớp có cấu trúc dạng cây, trong đó mỗi nút là một quyết định dựa trên một đặc trưng của dữ liệu, và mỗi nhánh tương ứng với kết quả của quyết định đó. Hai phương pháp phổ biến để xây dựng cây quyết định là **CART** và **ID3**.

**- Thuật toán CART (Classification and Regression Trees)**

**Khái niệm :** CART là thuật toán cây quyết định được sử dụng để phân lớp và hồi quy. Phân nhánh trong cây CART dựa trên chỉ số Gini nhằm tìm ra cách phân chia dữ liệu tốt nhất tại mỗi nút. Chỉ số Gini giúp chọn các đặc trưng sao cho độ hỗn loạn của lớp tại các nhánh con thấp nhất.

**Ưu điểm**

* **Đơn giản và dễ hiểu**: CART cung cấp một cây quyết định với các điều kiện rõ ràng.
* **Xử lý được dữ liệu số và phân loại**: Có thể xử lý cả dữ liệu định tính và định lượng.
* **Phân nhánh tốt khi dữ liệu đa dạng**: Chỉ số Gini giúp tối ưu hóa cây quyết định, đặc biệt hữu ích khi có nhiều lớp.

**Nhược điểm**

* **Dễ bị quá khớp**: CART có thể tạo ra cây phức tạp nếu không có biện pháp ngăn chặn quá khớp, chẳng hạn như giới hạn độ sâu của cây hoặc sử dụng cắt tỉa (pruning).
* **Kém chính xác với dữ liệu không cân bằng**: CART có thể ưu tiên các lớp phổ biến hơn trong quá trình phân lớp.
* **Khó mở rộng cho bài toán phức tạp**: Khi có quá nhiều đặc trưng hoặc dữ liệu lớn, CART có thể tạo ra cây lớn và khó diễn giải.
* **Mục tiêu**: CART xây dựng cây quyết định bằng cách phân chia dữ liệu dựa trên chỉ số **Gini**.
* **Gini Index**: Đây là thước đo để đánh giá độ hỗn loạn của các lớp trong một tập dữ liệu. Gini cho biết xác suất mà một điểm dữ liệu ngẫu nhiên sẽ bị phân loại sai nếu nó được gán nhãn dựa trên phân phối của lớp đó. Chỉ số Gini thấp tương đương với độ hỗn loạn thấp.
* **Quy trình**: CART chia dữ liệu tại mỗi nút dựa trên chỉ số Gini tối ưu, nghĩa là tại mỗi bước phân chia, nó chọn cách phân chia giúp giảm chỉ số Gini nhiều nhất, từ đó xây dựng cây cho đến khi các nút lá chứa các mẫu thuần nhất hoặc đạt ngưỡng dừng.

**-Thuật toán ID3 (Iterative Dichotomiser 3)**

**Khái niệm :** ID3 là thuật toán cây quyết định sử dụng Information Gain để phân chia dữ liệu. Information Gain giúp lựa chọn đặc trưng tốt nhất để phân chia dữ liệu bằng cách giảm entropy (mức độ hỗn loạn của thông tin).

**Ưu điểm**

* **Tốt khi dữ liệu có nhiều đặc trưng phân loại: Information Gain giúp ID3 xây dựng cây quyết định có khả năng phân biệt rõ ràng.**
* **Giải thích dễ dàng: Các điều kiện trong cây ID3 dựa trên Information Gain nên dễ diễn giải.**
* **Cấu trúc cây gọn: Information Gain thường giúp giảm số lượng đặc trưng không cần thiết, giúp cây ID3 gọn và dễ hiểu.**

**Nhược điểm**

* **Không tốt cho dữ liệu liên tục: ID3 không phù hợp với dữ liệu liên tục hoặc dữ liệu có giá trị số lớn, vì phải thực hiện quá trình rời rạc hóa dữ liệu.**
* **Dễ bị quá khớp: Giống như CART, ID3 có thể dễ dàng bị quá khớp khi dữ liệu huấn luyện có nhiễu hoặc không đủ tính đại diện.**
* **Độ phức tạp tăng với dữ liệu lớn: Khi số lượng đặc trưng hoặc mẫu lớn, thời gian xây dựng cây sẽ tăng đáng kể.**
* **Mục tiêu**: ID3 phân chia dữ liệu dựa trên **Information Gain**.
* **Information Gain**: Là thước đo mức độ giảm bất định về thông tin sau khi dữ liệu được phân chia bởi một đặc trưng. Nó tính toán sự khác biệt về **entropy** (mức độ hỗn loạn thông tin) giữa tập dữ liệu trước và sau khi phân chia. Đặc trưng nào tạo ra Information Gain lớn nhất sẽ được chọn để phân chia.
* **Quy trình**: Tại mỗi nút, ID3 chọn đặc trưng với Information Gain lớn nhất để phân chia dữ liệu, xây dựng cây sao cho các lớp trong các nút lá ngày càng thuần nhất.

**3. Đánh giá hiệu suất mô hình**

Sau khi mô hình cây quyết định được huấn luyện, hiệu suất của nó được đánh giá bằng cách dự đoán lớp cho tập kiểm thử và so sánh kết quả với nhãn thật. Một số chỉ số đánh giá thường dùng:

* **Độ chính xác (Accuracy)**: Tỷ lệ mẫu dự đoán đúng trên tổng số mẫu.
* **Độ nhạy (Recall)** và **Độ chính xác (Precision)**: Được sử dụng khi dữ liệu không cân bằng.
* **F1-score**: Trung bình điều hòa giữa Precision và Recall, phù hợp khi cần đánh giá hiệu suất tổng quát của mô hình.

**3. Kết luận**

Trong bài toán phân lớp ảnh sử dụng bộ dữ liệu IRIS và bộ 300 ảnh nha khoa, cả hai thuật toán **CART** và **ID3** đều có khả năng xây dựng cây quyết định hiệu quả với các đặc trưng được trích xuất từ ảnh.

* **CART** có lợi thế với dữ liệu có nhiều lớp và khả năng xử lý dữ liệu hỗn hợp, tuy nhiên, có thể cần áp dụng thêm cắt tỉa cây để tránh quá khớp.
* **ID3** cho thấy hiệu quả tốt khi dữ liệu có tính phân loại rõ ràng, nhưng cần chuẩn bị dữ liệu kỹ hơn đối với các đặc trưng liên tục và cần tối ưu hóa để giảm bớt nhiễu.

Vì vậy, trong trường hợp bộ dữ liệu ảnh đã được xử lý và trích xuất đặc trưng rõ ràng (đặc biệt là các đặc trưng định tính), ID3 có thể cho ra kết quả phân lớp tốt và cây đơn giản hơn. Tuy nhiên, nếu dữ liệu có nhiều nhiễu hoặc chứa các đặc trưng liên tục, **CART** có thể là lựa chọn tốt hơn nhờ tính linh hoạt và khả năng phân chia mạnh mẽ của nó.