**2.1. Phương hướng tiếp cận bài toán**

Quá trình giải quyết bài toán bắt đầu từ việc thu thập và tiền xử lý dữ liệu, khám phá và phân tích bộ dữ liệu thực nghiệm liên quan tới các đặc điểm và thu nhập của mỗi nhân viên. Sau đó tiến hành nghiên cứu và xây dựng mô hình và áp dụng để dự đoán tổng thu nhập của mỗi nhân viên dựa trên các đặc điểm. Cuối cùng là đánh giá mô hình, tinh chỉnh và cải thiện mô hình.

**2.2 Mô hình cây quyết định**

**2.2.1. Giới thiệu**

Cây quyết định là một phương pháp phân loại và dự đoán phổ biến trong lĩnh vực học máy (ML) và khai phá dữ liệu (DM). Thuật toán hoạt động bằng cách học các quy tắc quyết định đơn giản được suy ra từ các đặc điểm dữ liệu. Sau đó, quy tắc này có thể được sử dụng để dự đoán giá trị của mục tiêu cho mẫu dữ liệu mới.

Cây quyết định được biểu diễn dưới dạng cấu trúc cây, trong đó mỗi nút bên trong biểu diễn một thuộc tính, mỗi nhánh biểu diễn một quy tắc quyết định và mỗi nút lá biểu diễn một dự đoán. Thuật toán hoạt động bằng cách đệ quy chia dữ liệu thành các tập con ngày càng nhỏ hơn dựa trên các giá trị thuộc tính. Tại mỗi nút, thuật toán chọn thuộc tính chia dữ liệu tốt nhất thành các nhóm có giá trị mục tiêu khác nhau.

Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

Ta sẽ xét một ví dụ về cây quyết định. Giả sử dựa theo lương cơ bản, lương làm thêm và các khoản phúc lợi phụ cấp khác. Mục tiêu là dựa trên các đặc điểm này và dự đoán tổng thu nhập nhân viên của một mẫu dữ liệu mới?

A diagram of a flowchart

Description automatically generated

*Hình 2.1: Ví dụ về cây quyết định*

Dựa theo mô hình trên, ta có thể thấy rằng nếu nhân viên có lương cơ bản lớn hơn 16000$ và có lương làm thêm lớn hơn 30000$ thì nhân viên sẽ được có thu nhập cao hơn so với nhân viên có lương cơ bản bé hơn 16000$ và có lương làm thêm lớn hơn 10000$.

**2.2.2 Đặc điểm**

* **Các thành phần của cây quyết định**
* Node gốc: Là nút trên cùng trong cây, biểu diễn toàn bộ tập dữ liệu. Đây là khởi đầu của quá trình ra quyết định.
* Node nội bộ: Một nút tượng trưng cho sự lựa chọn liên quan đến một thuộc tính đầu vào. Phân nhánh các nút nội bộ kết nối chúng với các nút là hoặc các nút nội bộ khác.
* Node lá/ Node đầu cuối: Một nút không có bất kỳ nút con nào biểu thị nhãn lớp hoặc giá trị số.
* Node cha: Nút chia thành một hoặc nhiều nút con.
* Node con: Các nút xuất hiện khi nút cha bị tách
* **Thuật toán ID3**
* Thuật toán ID3 là một trong những thuật toán đầu tiên và được sử dụng nhiều nhất, được Ross Quinlan tạo ra vào năm 1986. Thuật toán ID3 xây dựng một cây quyết định từ một tập dữ liệu nhất định bằng phương pháp tham lam, từ trên xuống dưới.
* Thuật toán hoạt động bằng cách sử dụng phương pháp tham lam chọn thuộc tính tối đa hóa mức tăng thông tin tại mỗi node. ID3 tính toán Entropy và mức tăng thông tin cho mỗi thuộc tính và chọn thuộc tính có mức tăng thông tin cao nhất để phân tách.
* Thuật toán ID3 sử dụng Entropy để đo lường sự không chắc chắn trong một tập dữ liệu và Information Gain để đánh giá và lựa chọn các thuộc tính từ đó tối ưu hóa việc phân chia dữ liệu và xây dựng mô hình cây quyết định hiệu quả hơn, dẫn đến những dự đoán chính xác hơn.

Entropy (H(D)) là thước đo mức độ tinh khiết trong tập dữ liệu, được tính theo công thức:

H(D)-i=1c Pi log2 Pi

Trong đó:

* H(D): Entropy của tập dữ liệu D.
* c: Số lượng lớp (Classes) trong tập dữ liệu.
* Pi: Xác suất của lớp i trong tập dữ liệu D, được tính bằng cách chia số lượng mẫu trong lớp i cho tổng số mẫu trong D.

Information gain (IG(A)) là độ thu lợi về thông tin, độ quan trọng về 1 thuộc tính về mặt thông tin, được tính theo công thức:

Gain (S, A) = H(S) - H(S|A)

Trong đó:

* H(S): Entropy của tập dữ liệu S trước khi phân chia
* H(S|A): Entropy của tập dữ liệu S sau khi phân chia theo thuộc tính A.
* **Thuật toán C4.5**

 Thuật toán C4.5 là thuật toán được Ross Quinlan cải tiến từ thuật toán ID3. Trong các ứng dụng học máy và khai thác dữ liệu, đây là một phương pháp được ưa chuộng để tạo cây quyết định. C4.5 được tạo ra để khắc phục một số nhược điểm của ID3, bao gồm khả năng xử lý các điểm liên tục, thực hiện cắt tỉa cây để giảm xu hướng overfitting với tập huấn luyện và sử dụng tỉ lệ tăng thông tin (Information gain ratio) thay vì tăng thông tin (Information gain) như ID3, giúp giảm thiểu thiên lệch với thuộc tính có nhiều giá trị.

* **Một số thuật toán khác**

Bên cạnh ID3, C4.5, mô hình cây quyết định còn một số thuật toán khác như:

* Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng kiểm định Chi-square để xác định phân chia tốt nhất, phù hợp nhất cho các dữ liệu có phân phối không đồng đều hoặc không liên tục.
* Thuật toán CART: sử dụng tạp chất Gini để phân loại. Khi chọn một thuộc tính để phân chia, thuật toán sẽ tính toán tạp chất Gini cho mỗi phân chia có thể và chọn phân chia có tạp chất thấp nhất.
* MARS
* Conditional Inference Trees

**2.2.3 Ưu điểm và nhược điểm**

Cây quyết định là một thuật toán phổ biến trong lĩnh vực học máy và khai thác dữ liệu, được sử dụng rộng rãi cho các bài toán phân loại và dự đoán. Cây quyết định cũng có những ưu điểm và nhược điểm cần cân nhắc trước khi lựa chọn và triển khai:

* Ưu điểm:
* Mô hình dễ dàng triển khai và có thể xây dựng nhanh chóng ngay cả với những người không chuyên sâu về học máy.
* Mô hình sinh ra các quy tắc dễ hiểu cho người sử dụng và người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
* Có thể xử lý cả dữ liệu phân loại và dữ liệu liên tục, linh hoạt trong nhiều ứng dụng khác nhau.
* Dễ dàng nắm bắt các quan hệ phi tuyến tính giữa các biến trong dữ liệu.
* Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
* Có khả năng làm việc với dữ liệu lớn.
* Nhược điểm:
* Cây quyết định có xu hướng tạo ra cây rất phức tạp khi cố gắng phân chia để đạt độ chính xác tối đa, dẫn đến overfitting, khiến mô hình không tổng quát tốt khi áp dụng trên dữ liệu mới.
* Không ổn định và nhạy cảm với các thay đổi của dữ liệu, dẫn đến sự thay đổi đáng kể trong cấu trúc cây.
* Tốn kém khi tính toán với dữ liệu lớn.