# Giới thiệu bài toán

# Giải quyết bài toán

## Cây quyết định (Decision tree)

### Giới thiệu về cây quyết định

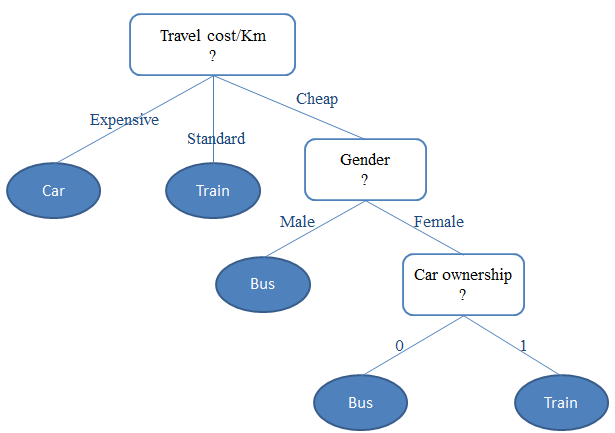
* Cây quyết định là một cây phân cấp có cấu trúc.
* Dùng để phân lớp đối tượng dựa vào dãy các luật (series of rules), các luật này được sinh ra từ tập dữ liệu (training set).
* Các thuộc tính phân lớp có kiểu dữ liệu là binary, nominal, ordinal, continuos.
* Ví dụ: Cho tập dữ liệu (training set) như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thuộc tính (Attribute) | | | | Thuộc tính phân lớp (Class) |
| Gender | Car ownership | Travel cost ($/km) | Income level | Transportation mode |
| Male | 0 | Cheap | Low | Bus |
| Male | 1 | Cheap | Medium | Bus |
| Female | 1 | Cheap | Medium | Train |
| Female | 0 | Cheap | Low | Bus |
| Male | 1 | Cheap | Medium | Bus |
| Male | 0 | Standard | Medium | Train |
| Female | 1 | Standard | Medium | Train |
| Female | 1 | Expensive | High | Car |
| Male | 2 | Expensive | Medium | Car |
| Female | 2 | Expensive | High | Car |

Bảng 1. Ví dụ bảng dữ liệu lựa chọn phương tiện di chuyển

Tập dữ liệu trên mô tả 10 đối tượng, mỗi đối tượng được miêu tả bằng 4 thuộc tính là “Gender” (kiểu dữ liệu binary), “Car ownership” (quantitative interger), “Travel cost/km” (ordinal), “Income level” (ordinal) và thuộc tính phân loại – category attribute – “Transportation mode” (ordinal).

Từ tập dữ liệu trên, chúng ta có thể tạo ra cây quyết định như sau:



Hình 1. Ví dụ về cây quyết định

Trong cây quyết định trên, thuộc tính “Income level” không xuất hiện trong cây; vì dựa vào bảng 1 (trang 1), thuộc tính “Travel cost/Km” sẽ sinh ra cây quyết định dùng để phân lớp tốt hơn “Income level”

### Sử dụng cây quyết định để dự đoán lớp các dữ liệu chưa biết

* Mục đích chính của cây quyết định là dùng để xác định lớp hay nói khác đi là dự đoán lớp của các dữ liệu chưa biết dựa vào cây quyết định được sinh ra từ tập dữ liệu (training data)
* Ví dụ: dựa vào ví dụ ở phần 2.1.1 Giới thiệu về cây quyết định (trang 1). Cho tập dữ liệu cần dự đoán sau đây:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Gender | Car ownership | Travel cost ($/km) | Income level | Transportation mode |
| Ngoc | Female | 1 | Cheap | High | ? |
| Hieu | Male | 0 | Standard | High | ? |
| Phuc | Male | 1 | Cheap | Medium | ? |
| Nguyen | Male | 2 | Expensive | High | ? |

Bảng 2. Ví dụ về bảng dữ liệu cần dự đoán phương tiện di chuyển 1

Dựa vào cây quyết định (Hình 1, trang 2), cây quyết định sẽ được duyệt từ nút gốc “Travel cost/km”, dãy các luật sau sẽ được sinh ra:

* Nếu “Travel cost/Km” là **Expensive** thì người đó sẽ chọn phương tiện di chuyển là **car**.
* Nếu “Travel cost/Km” là **Standard** thì người đó sẽ chọn phương tiện di chuyển là **train**
* Nếu “Travel cost/Km” là **Cheap**, chúng ta sẽ xem xét thuộc tính “Gender”:
  + Nếu “Gender” là **Male**, người đó sẽ chọn phương tiện là **bus**
  + Nếu “Gender là **Female**, thì xem người đó sở hữu bao nhiêu xe hơi (thuộc tính “Car ownership”). Nếu số xe sở hữu là **0**, thì người đó sẽ chọn phương tiện di chuyển là **bus**; ngược lại, nếu số xe sở hữu lớn hơn hay bằng **1**, thì người đó sẽ chọn phương tiện di chuyển là **train**.

Bảng 2 sẽ được dự đoán như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Gender | Car ownership | Travel cost ($/km) | Income level | Transportation mode |
| Ngoc | Female | 1 | Cheap | High | Train |
| Hieu | Male | 0 | Standard | High | Train |
| Phuc | Male | 1 | Cheap | Medium | Bus |
| Nguyen | Male | 2 | Expensive | High | Car |

Bảng 3. Ví dụ về bảng dữ liệu cần dự đoán phương tiện di chuyển 2

* **Một số chú ý khi sử dụng cây quyết định:**
* Phụ thuộc rất nhiều vào training data, tập dữ liệu training data càng lớn thì cây quyết định sẽ đáng tin cậy hơn.
* Không thể nói cây quyết định được sinh ra từ cây quyết định trên là tập luật tốt nhất
* Có nhiều thuật toán phân lớp: ID3, J48, C4.5, C5, CART (Classification and Regression Tree), … Việc lựa chọn thuật toán phụ thuốc vào rất nhiều yếu tố, trong đó yếu tố cấu trúc dữ liệu ảnh hưởng rất nhiều đến kết quả của thuật toán. Chẳn hạn, thuật toán ID3 và CART hiệu quả cho việc phân lớp đối với các dữ liệu số (quantitative value), trong khi đó, thuật toán J48, C4.5 có hiệu quả hơn đối với dữ liệu Qualititive value (ordinal, binary, nominal)

### Đặc điểm của cây quyết định

### Một số độ đo thông dụng

Cho bảng dữ liệu bao gồm các thuộc tính và thuộc tính phân lớp, chúng ta có thể đo được tính đồng nhất hay không đồng nhất thông qua thuộc tính phân lớp. Bảng dữ liệu có tính đồng nhất nếu nó có duy nhất một phân lớp. Ngược lại, nếu nó có nhiều phân lớp khác nhau, thì bảng dữ liệu có tính không đồng nhất hay tính pha trộn. Chúng ta có thể đo được mức độ pha trộn (Impurity Degree); Entropy, độ đo Gini và classification error là những cách tính mức độ pha trộn thông dụng nhất.

Pj: xác suất xảy ra phân lớp j

Ví dụ: Chúng ta cùng nhìn lại Bảng 1. Ví dụ bảng dữ liệu lựa chọn phương tiện di chuyển. “Transportation mode” có 3 nhóm Bus, Car và Train. Bảng 1 có 10 dòng dữ liệu, trong đó “Transportation mode” có 4 buses, 3 cars, 3 trains (4B, 3C, 3T). Xác suất để xảy ra các phân lớp là:

**Entropy**

Khi bảng dữ liệu có duy nhất 1 phân lớp thì Entropy sẽ bằng 0 vì xác suất là 1 và . Entropy sẽ đạt giá trị cực đại khi tất cả thuộc tính phân lớp có xác suất xảy ra bằng nhau. Biểu đồ dưới đây thể hiện giá trị cực đại của Entropy sẽ thay đổi phụ thuộc vào số lượng thuộc tính phân lớp n, trong trường hợp xác xuất tất cả thuộc tính phân lớp thì . Giá trị của Entropy sẽ lớn hơn 1 khi số lượng thuộc tính phân lớp nhiều hơn 2



Biểu đồ 1. Mối liên hệ giữa giá trị Entropy cực đại và số lượng phân lớp

**Gini index**

Một cách khác để tính “Impurity degree”.

Chúng ta cùng quay lại ví dụ ở trên

Khi bảng dữ liệu có duy nhất 1 phân lớp thì Gini index sẽ bằng 0 vì xác suất bằng 1 và . Cũng giống như Entropy, Gini index sẽ đạt giá trí cực đại khi tất cả thuộc tính phân lớp có xác suất xảy ra bằng nhau. Biểu đồ dưới đây thể hiển giá trị cực đại của Gini index sẽ khác nhau phụ thuộc vào số lượng phân lớp n, khi xác xuất tất cả thuộc tính phân lớp



Biểu đồ 2. Mối liên hệ giữa giá trị Gini index cực đại và số lượng phân lớp

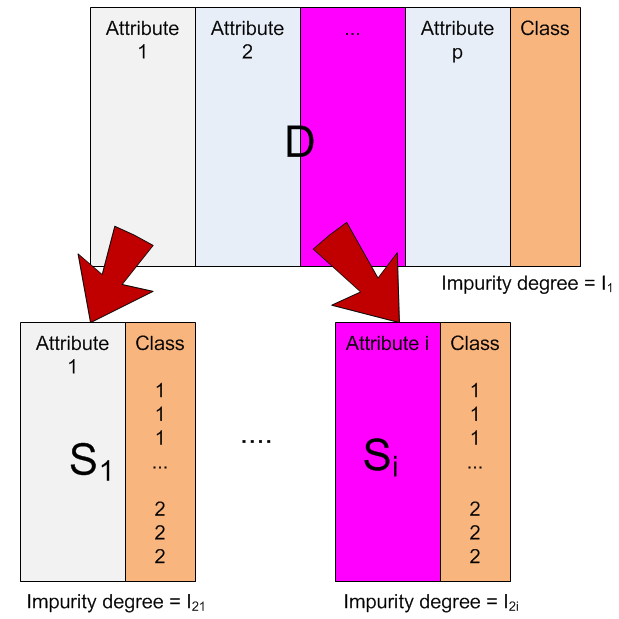
**Classification error**

Trong ví dụ trên

Cũng giống như Entropy và Gini index, Classification error sẽ bằng 0 khi bảng dữ liệu có duy nhất 1 phân lớp vì xác suất bằng 1 và .

### Nguyên tắc hoạt động của một số thuật toán xây dựng cây quyết định thường được sử dụng

* Các thuật toán xây dựng cây quyết định thường được sử dụng nhất là ID3, C4.5 và CART (classification and regression trees).
* Nhìn chung, các thuật toán xây dựng cây quyết định đều được xây dựng trên quy tắc đệ quy. Ví dụ, Hunt là thuật toán dựng cây quyết định, đệ quy theo nút của cây, bắt đầu từ nút gốc. Mặc dù kết quả đạt được từ thuật toán Hunt không được tối ưu, nhưng đây là một trong những một trong những thuật toán dùng để xây dựng cây quyết định sớm nhất.
* Giả sử, chúng ta có một bảng dữ liệu chứa các thuộc tính và thuộc tính phân lớp, tạm gọi bảng này là D. Từ bảng D, chúng ta sẽ lấy ra từng cột thuộc tính trong bảng để đối chiếu với các giá trị của thuộc tính phân lớp. Nếu chúng ta có p cột dữ liệu, và lấy từng phần tử p là tập con của D, tạm gọi từng tập hợp này là Si. Bảng D là tập hợp những Si và thuộc tính phân lớp.



* Sau khi phân chia bảng dữ liệu D như trên, chúng ta sẽ tính mức độ pha trộn (tham khảo cách tính tại phần 2.1.4 Một số độ đo thông dụng)
* Ví dụ, dựa vào Bảng 1. Ví dụ bảng dữ liệu lựa chọn phương tiện di chuyển. chúng ta có thể tính “Impurity degree” dựa vào thuộc tính phân lớp “Transportation mode”. “Transportation mode” có 4 busses, 3 cars và 3 trains (4B, 3C, 3T)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thuộc tính (Attribute) | | | | Thuộc tính phân lớp (Class) |
| Gender | Car ownership | Travel cost ($/km) | Income level | Transportation mode |
| Male | 0 | Cheap | Low | Bus |
| Male | 1 | Cheap | Medium | Bus |
| Female | 1 | Cheap | Medium | Train |
| Female | 0 | Cheap | Low | Bus |
| Male | 1 | Cheap | Medium | Bus |
| Male | 0 | Standard | Medium | Train |
| Female | 1 | Standard | Medium | Train |
| Female | 1 | Expensive | High | Car |
| Male | 2 | Expensive | Medium | Car |
| Female | 2 | Expensive | High | Car |

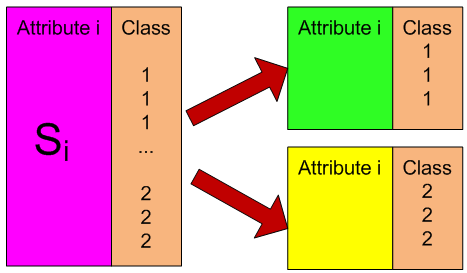
4B, 3C, 3T

Entropy 1.571

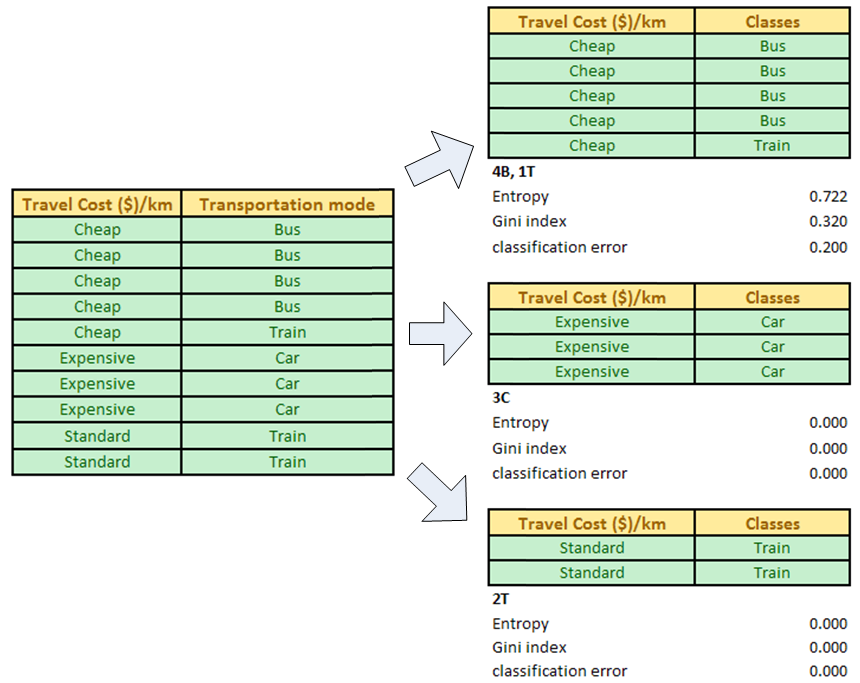
Gini index 0.660

Classification index 0.600

* Sau khi phân chia bảng D thành từng tập con Si, chúng ta tiếp tục phân chia Si để tính “Impurity degree” cho từng giá trị thuộc tính trong từng tập Si



* Tiếp tục ví dụ trên, thuộc tính “Travel cost/km” có 3 giá trị, đó là “Cheap”, “Standard” và “Expensive”.



**Độ lợi thông tin (Information Gain):**

### So sánh một số thuật toán xây dựng cây quyết định

### Liên kết các cây quyết định

### Ứng dụng cây quyết định vào bài toán