TRƯỜNG ĐẠI HỌC KHOA HỌC
KHOA CÔNG NGHỆ THÔNG TIN

HỌ VÀ TÊN SINH VIÊN: PHAN MINH PHƯỚC
MÃ SINH VIÊN: 18T1021245

TÊN HỌC PHẦN : THỰC TẬP VIẾT NIÊN LUẬN

ĐỀ TÀI: Tìm hiểu thuật toán ID3 trong xây dựng cây quyết định và khai thác bằng phần mềm WEKA để phân lớp dữ liệu

GIẢNG VIÊN HƯỚNG DẪN: LÊ MẠNH THẠNH

HUẾ, NĂM 2021

TRƯỜNG ĐẠI HỌC KHOA HỌC
KHOA CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

PHIẾU ĐÁNH GIÁ

Học kỳ 2 Năm học 2020-2021

Cán bộ chấm thi 1	**Cán bộ chấm thi 2**
Nhận xét:	**Nhận xét**: .....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....
Điểm đánh giá của CBChT1:	**Điểm đánh giá của CBChT2:**
Bằng số:	Bằng số: .....
Bằng chữ:	Bằng chữ: .....

Điểm kết luận: Bằng số..... Bằng chữ:.....

Thừa Thiên Huế, ngày …… tháng …… năm 20…

CBChT2

(Ký và ghi rõ họ tên)

(Ký và ghi rõ họ tên)

Tìm hiểu thuật toán ID3 trong xây dựng cây quyết định và khai thác bằng phần mềm WEKA để phân lớp dữ liệu.

I. Tổng quan về khai phá dữ liệu:

1.1. Tại sao lại cần khai phá dữ liệu ?

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử (đĩa cứng, CD -ROM, băng từ, .v.v.) không ngừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số lượng cũng như kích cỡ của các cơ sở dữ liệu (CSDL) cũng tăng lên một cách nhanh chóng. Nói một cách hình ảnh là chúng ta đang "ngập" trong dữ liệu nhưng lại "đói" tri thức. Câu hỏi đặt ra là liệu chúng ta có thể khai thác được gì từ những "núi" dữ liệu tưởng chừng như "bỏ đi" ấy không ?

"Necessity is the mother of invention"- Data Mining ra đời như một hướng giải quyết hữu hiệu cho câu hỏi vừa đặt ra ở trên []. Khá nhiều định nghĩa về Data Mining và sẽ được đề cập ở phần sau, tuy nhiên có thể tạm hiểu rằng Data Mining như là một công nghệ tri thức giúp khai thác những thông tin hữu ích từ những kho dữ liệu được tích trữ trong suốt quá trình hoạt động của một công ty, tổ chức nào đó.

1.2. Khai phá dữ liệu là gì ?

Định nghĩa: Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó.

1.3. Các chức năng của khai phá dữ liệu:

Data Mining được chia nhỏ thành một số hướng chính như sau:

- Mô tả khái niệm (concept description): thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.
- Luật kết hợp (association rules): là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: "60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô". Luật kết hợp được ứng dụng nhiều trong lĩnh vực kính doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, .v.v.
- Phân lớp và dự đoán (classification & prediction): xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), .v.v. Người ta còn gọi phân lớp là học có giám sát (học có thầy).
- Phân cụm (clustering): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát (học không thầy).

• Khai phá chuỗi (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

1.4. Ứng dụng của khai phá dữ liệu :

- Phân tích thị trường và chứng khoán
- Phát hiện gian lận
- Quản lý rủi ro và phân tích doanh nghiệp
- Phân tích giá trị trọn đời của khách hàng
- Điều trị y học (medical treatment)
- Bảo hiểm (insurance)
- Nhận dạng (pattern recognition)

1.5. Các kỹ thuật trong khai phá dữ liệu:

- **Kỹ thuật phân tích phân loại (Classification Analysis)**

Kỹ thuật khai phá dữ liệu đầu tiên là kỹ thuật phân tích phân loại. Đây là kỹ thuật cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước.

Bạn có thể sử dụng kỹ thuật này để phân loại khách hàng, mặt hàng, v.v bằng cách mô tả nhiều thuộc tính để phân loại đối tượng vào một lớp cụ thể.

Chúng ta thường sử dụng kỹ thuật khai thác dữ liệu này để lấy các thông tin quan trọng từ dữ liệu và siêu dữ liệu. Vì vậy, trong phân tích phân loại, chúng ta cần áp dụng các thuật toán khác nhau tùy thuộc vào mục tiêu sử dụng.

Ví dụ, Email Outlook sử dụng các thuật toán nhất định để mô tả một email là hợp pháp hoặc spam. Hay các doanh nghiệp có thể áp dụng kỹ thuật này để phân loại khách hàng theo đối tượng hay độ tuổi.

- **Kỹ thuật Association Rule Learning :**

Kỹ thuật Association Rule Learning trong khai phá dữ liệu được sử dụng để xác định mối quan hệ giữa các biến khác nhau trong cơ sở dữ liệu. Ngoài ra, nó còn được sử dụng để "giải nén" các mẫu ẩn trong dữ liệu. Association Rule rất hữu ích để kiểm tra, dự đoán hành vi và thường được áp dụng trong ngành bán lẻ.

Thêm vào đó, các doanh nghiệp sử dụng kỹ thuật này để xác định hành vi mua sắm, phân tích dữ liệu trong giỏ hàng của khách hàng tiềm năng. Trong lĩnh vực Công nghệ Thông tin, các lập trình viên sử dụng kỹ thuật này để xây dựng các chương trình Machine Learning.

- **Kỹ thuật phát hiện bất thường (Anomaly or Outlier Detection):**

Về cơ bản, kỹ thuật khai phá dữ liệu (Data Mining) này dùng để nhấn mạnh vào việc quan sát các mục dữ liệu trong bộ dữ liệu để tìm ra các tập dữ liệu không khớp

với mẫu dự kiến. Bất thường ở đây có thể đề cập đến độ lệch, sự khác thường, các nhiễu và ngoại lệ.

Sự bất thường được xem là khá quan trọng vì nó có thể cung cấp một số thông tin cần thiết. Nó có thể là một dữ liệu khác biệt so với mức trung bình chung trong một tập dữ liệu. Điều này chỉ ra rằng một cái gì đó khác thường đã xảy ra và các nhà phân tích dữ liệu cần chú ý.

Kỹ thuật này có thể được sử dụng trong nhiều lĩnh vực khác nhau. Chẳng hạn như phát hiện xâm nhập hay theo dõi sức khỏe.

- **Kỹ thuật phân tích theo cụm (Clustering Analysis):**

"Cụm" có nghĩa là một nhóm các đối tượng dữ liệu. Các đối tượng tương tự nhau thì sẽ nằm trong một cụm. Kết quả là các đối tượng tương tự nhau trong cùng một nhóm. Về cơ bản, kỹ thuật khai phá dữ liệu này thường được ứng dụng để tạo hồ sơ khách hàng. Hoặc trong lĩnh vực Marketing, đây được xem là việc chia phân khúc khách hàng.

- **Kỹ thuật phân tích hồi quy (regression analysis):**

Theo thuật ngữ thống kê, phân tích hồi quy được sử dụng để xác định và phân tích mối quan hệ giữa các biến. Nó giúp bạn hiểu giá trị đặc trưng của sự thay đổi ở các biến phụ thuộc.

- **Kỹ thuật dự báo (prediction):**

Trong khai phá dữ liệu, kỹ thuật dự báo được ứng dụng ở một số trường hợp đặc biệt. Nó được sử dụng để khám phá mối quan hệ giữa các biến độc lập và phụ thuộc. Chẳng hạn, bạn có thể sử dụng kỹ thuật dự báo cho việc bán hàng để dự đoán lợi nhuận cho tương lai. Giả sử, bán hàng là một biến độc lập, lợi nhuận có thể là một biến phụ thuộc. Khi đó, chúng ta có thể vẽ đường cong hồi quy để dự đoán lợi nhuận.

II. Cây quyết định :

Khái niệm Cây quyết định **(Decision Tree)**

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

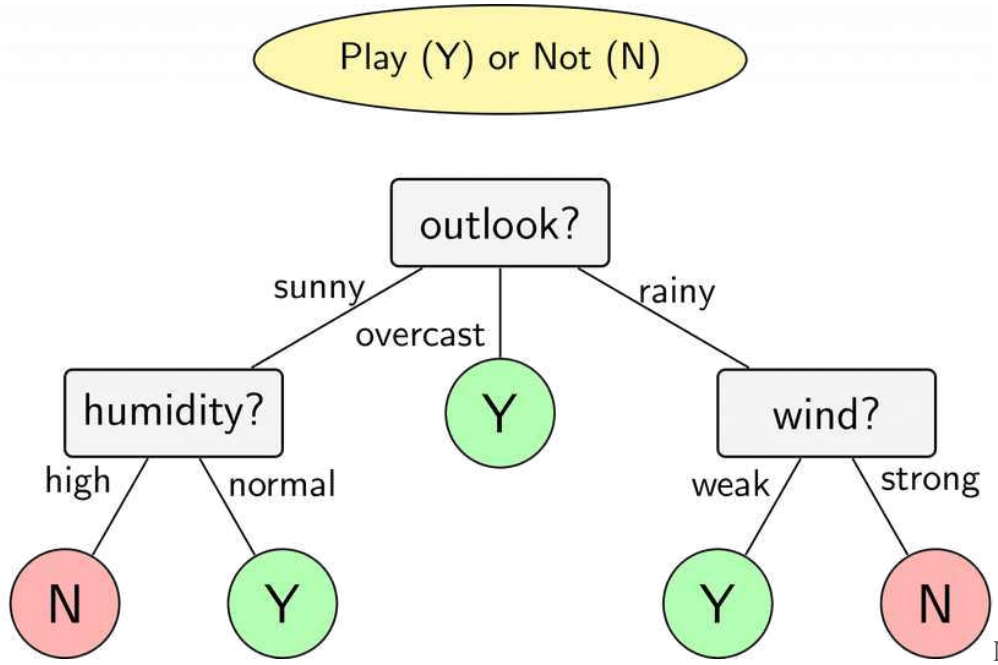
Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

Ta hãy xét một ví dụ 1 kinh điển khác về cây quyết định. Giả sử dựa theo thời tiết mà các bạn nam sẽ quyết định đi đá bóng hay không?

Những đặc điểm ban đầu là:

- Thời tiết
- Độ ẩm
- Gió

Dựa vào những thông tin trên, bạn có thể xây dựng được mô hình như sau:



Mô hình cây quyết định

Dựa theo mô hình trên, ta thấy:

Nếu trời nắng, độ ẩm bình thường thì khả năng các bạn nam đi chơi bóng sẽ cao. Còn nếu trời nắng, độ ẩm cao thì khả năng các bạn nam sẽ không đi chơi bóng.

Entropy trong Cây quyết định (Decision Tree)

Entropylà thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với công thức như sau:

Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau $x_1, x_2, \ldots, x_n$.

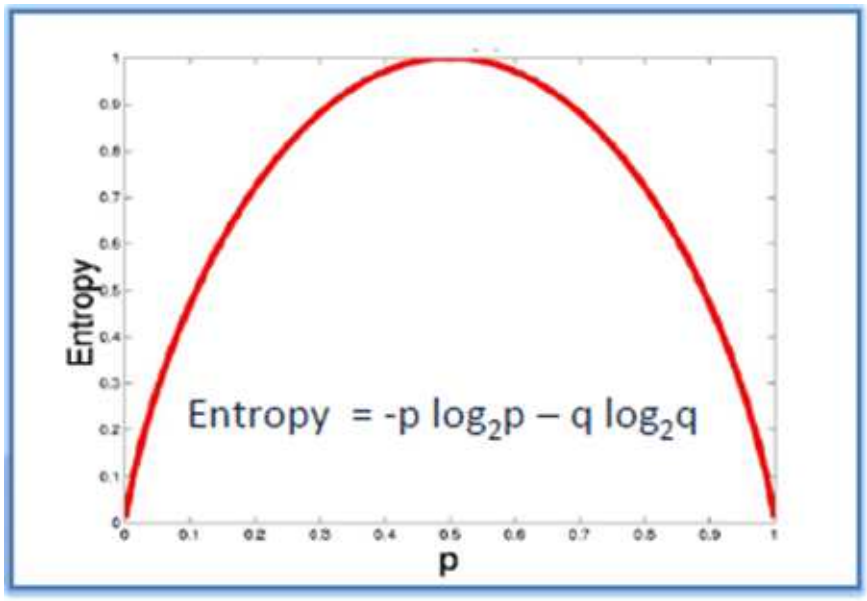
Giả sử rằng xác suất để x nhận các giá trị này là $p_i = p(x = x_i)$.

Ký hiệu phân phối này là $p=(p_1, p_2, \ldots, p_n)$. Entropy của phân phối này được định nghĩa là:

$$H(p) = -\sum_{n=1}^{n} p_i \log(p_i)$$

Giả sử bạn tung một đồng xu, Entropy sẽ được tính như sau:

$$H = -[0.5 \ln(0.5) + 0.5 \ln(0.5)]$$



$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

Hàm Entropy

Hình vẽ trên biểu diễn sự thay đổi của hàm Entropy. Ta có thể thấy rằng, Entropyđạt tối đa khi xác suất xảy ra của hai lớp bằng nhau.

- P tinh khiết: $p_i = 0$ hoặc $p_i = 1$
- P vẩn đục: $p_i = 0.5$, khi đó hàm Entropy đạt đỉnh cao nhất

Information Gain trong Cây quyết định (Decision Tree)

Information Gain dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất.

Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:

•**Bước 1**: Tính toán hệ số Entropycủa biến mục tiêu S có N phần tử với N_c phần tử thuộc lớp c cho trước:

$$H(S)= -\sum_{c=1}^{c} (N_c/N) \log(N_c/N)$$

•**Bước 2**: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S_1, S_2, …, S_K với số điểm trong mỗi child node lần lượt là m_1, m_2 ,…, m_K , ta có:

$$H(x, S) = \sum_{k=1}^{K} (m_k / N) * H(S_k)$$

Bước 3: Chỉ số Gain Information được tính bằng:

$$\mathbf{G(x, S) = H(S) - H(x,S)}$$

Với ví dụ 2 trên, ta tính được hệ số Entropy như sau:

$Entropy_{Parent} = -(0.57*ln(0.57) + 0.43*ln(0.43))$ = 0.68

Hệ số Entropytheo phương pháp chia thứ nhất:

$Entropy_{left} = -(.75*ln(0.75) + 0.25*ln(0.25)) = 0.56$
$Entropy_{right} = -(.33*ln(0.33) + 0.67*ln(0.67)) = 0.63$

Ta có thể tính hệ số ***Information Gain*** như sau:

*Information Gain = 0.68 – (4*0.56 + 3*0.63)/7 = 0.09*

Hệ số Entropyvới phương pháp chia thứ hai như sau:

$Entropy_{left} = -(.67*ln(0.67) + 0.33*ln(0.33)) = 0.63$
$Entropy_{middle} = -(.5*ln(0.5) + 0.5*ln(0.5)) = 0.69$
$Entropy_{right} = -(.5*ln(0.5) + 0.5*ln(0.5)) = 0.69$

Hệ số ***Information Gain***:

*Information Gain = 0.68 – (3*0.63 + 2*0.69 + 2*0.69)/7 = 0.02*

So sánh kết quả, ta thấy nếu chia theo phương pháp 1 thì ta được giá trị hệ số Information Gain lớn hơn gấp 4 lần so với phương pháp 2. Như vậy, giá trị thông tin ta thu được theo phương pháp 1 cũng nhiều hơn phương pháp 2.

Thuật toán C4.5

Thuật toán C4.5 là thuật toán cải tiến của ID3.

Trong thuật toán ID3, Information Gain được sử dụng làm độ đo. Tuy nhiên, phương pháp này lại ưu tiên những thuộc tính có số lượng lớn các giá trị mà ít xét tới những giá trị nhỏ hơn. Do vậy, để khắc phục nhược điểm trên, ta sử dụng độ đo Gain Ratio (trong thuật toán C4.5) như sau:

Đầu tiên, ta chuẩn hoá information gain với trị thông tin phân tách (split information):

$$Gain\ Ratio = \frac{Information\ Gain}{Split\ Info}$$

Trong đó: Split Info được tính như sau:

$$-\sum_{i=1}^{n} Di\ log_2 Di$$

Giả sử chúng ta phân chia biến thành n nút cón và Di đại diện cho số lượng bản ghi thuộc nút đó. Do đó, hệ số Gain Ratio sẽ xem xét được xu hướng phân phối khi chia cây.

Áp dụng cho ví dụ trên và với cách chia thứ nhất, ta có

*Split Info = – ((4/7)*log₂(4/7)) – ((3/7)*log₂(3/7)) = 0.98*

Gain Ratio = 0.09/0.98 = 0.092

Tiêu chuẩn dừng

Trong các thuật toán Decision tree, với phương pháp chia trên, ta sẽ chia mãi các node nếu nó chưa tinh khiết. Như vậy, ta sẽ thu được một tree mà mọi điểm trong tập huấn luyện đều được dự đoán đúng (giả sử rằng không có hai input giống nhau nào cho output khác nhau). Khi đó, cây có thể sẽ rất phức tạp (nhiều node) với nhiều leaf node chỉ có một vài điểm dữ liệu. Như vậy, nhiều khả năng overfitting sẽ xảy ra.

Để tránh trường hợp này, ta có thể dừng cây theo một số phương pháp sau đây:

- nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một class.
- nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.
  nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế *chiều sâu của tree* này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.
- nếu tổng số leaf node vượt quá một ngưỡng nào đó.
- nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

Ngoài ra, ta còn có phương pháp cắt tỉa cây.

Một số thuật toán khác

Ngoài ID3, C4.5, ta còn một số thuật toán khác như:

- Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng thống kê chi-square để xác định các phân tách tối ưu. Các biến mục tiêu đầu vào có thể là số (liên tục) hoặc phân loại.
- Thuật toán C&R: sử dụng phân vùng đệ quy để chia cây. Tham biến mục tiêu có thể dạng số hoặc phân loại.
  MARS
- Conditional Inference Trees

Ưu/nhược điểm của thuật toán cây quyết định

Ưu điểm

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bởi những lợi ích của nó:

- Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
- Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
- Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
- Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
- Có khả năng là việc với dữ liệu lớn

Nhược điểm

Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

- Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
- Cây quyết định hay gặp vấn đề overfitting

III. Thuật toán ID3:

Giải thuật ID3 (gọi tắt là ID3) Được phát triển đồng thời bởi Quinlan trong AI và Breiman, Friedman, Olsen và Stone trong thống kê. ID3 là một giải thuật học đơn giản nhưng tỏ ra thành công trong nhiều lĩnh vực. ID3 là một giải thuật hay vì cách biểu diễn tri thức học được của nó, tiếp cận của nó trong việc quản lý tính phức tạp, heuristic của nó dùng cho việc chọn lựa các khái niệm ứng viên, và tiềm năng của nó đối với việc xử lý dữ liệu nhiễu.

ID3 biểu diễn các khái niệm (concept) ở dạng các cây quyết định (decision tree). Biểu diễn này cho phép chúng ta xác định phân loại của một đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó.

Như vậy, nhiệm vụ của giải thuật ID3 là học cây quyết định từ một tập các ví dụ rèn luyện (training example) hay còn gọi là dữ liệu rèn luyện (training data).

Input: Một tập hợp các ví dụ. Mỗi ví dụ bao gồm các thuộc tính mô tả một tình huống, hay một đối tượng nào đó, và một giá trị phân loại của nó.

Output: Cây quyết định có khả năng phân loại đúng đắn các ví dụ trong tập dữ liệu rèn luyện, và hy vọng là phân loại đúng cho cả các ví dụ chưa gặp trong tương lai.

Giải thuật ID3 xây dựng cây quyết định được trình bày như sau:

Lặp:
1. Chọn A <= thuộc tính quyết định “tốt nhất” cho nút kế tiếp
2. Gán A là thuộc tính quyết định cho nút
3. Với mỗi giá trị của A, tạo nhánh con mới của nút
4. Phân loại các mẫu huấn luyện cho các nút lá
5. Nếu các mẫu huấn luyện được phân loại hoàn toàn thì NGƯNG,
Ngược lại, lặp với các nút lá mới.

Thuộc tính tốt nhất ở đây là thuộc tính có entropy trung bình thấp nhất theo thuộc tính kết quả với Entropy được tính như sau:

- Gọi S là tập các mẫu huấn luyện
- Gọi p là tỷ lệ các mẫu dương trong S

  Ta có $H \equiv -p.\log_2 p - (1-p).\log_2(1-p)$

Entropy trung bình của một thuộc tính bằng trung bình theo tỉ lệ của entropy các nhánh:

$$AE(ĐHLTB) = \sum_{v \in Value(A)} p_v H_{Av}$$

Phân tích bài toán chơi golf

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

- Phân tích bài toán :

+ ta có :

* S= 14

*m=2 (2 kết quả)

*C1=”play”,C2=”no”.

*S

S1:Tổng các trường hợp của C_1:9

S2:Tổng các trường hợp của C_2:5

$I(S_1,S_2)=I(9,5)= -\frac{9}{14}\log_2\frac{9}{14} -\frac{5}{14}\log_2\frac{5}{14} = 0.940$

Độ lợi thông tin thuộc tín windy:

$$-\frac{3}{6}\log_2\frac{3}{6}-\frac{3}{6}\log_2\frac{3}{6}=1.000$$

$$-\frac{6}{8}\log_2\frac{6}{8}-\frac{2}{8}\log_2\frac{2}{8}=0.811$$

E(Windy) = 1 * 6/14 + 0.811 * 8/14 = 0.892

Gain(S, Windy) = 0.940 - 0.892 = 0.048

Độ lợi thông tin thuộc tính Humidity:

$$-\frac{3}{4}log_2\frac{3}{4}-\frac{1}{4}log_2\frac{1}{4}= \quad 0.811$$

$$-\frac{6}{10}log_2\frac{6}{10}-\frac{4}{10}log_2\frac{4}{10}= \quad 0.97$$

E(Humidity)= 0.811*4/14 + 0.97 * 10/14 = 0.924

Gain(S,Humidity) = 0.940-0.924=0.016

Độ lợi thông tin của thuộc tính Outlook:

$$-\frac{2}{5}log_2\frac{2}{5}-\frac{3}{5}log_2\frac{3}{5}=0.971$$

$$-\frac{3}{5}log_2\frac{3}{5}-\frac{2}{5}log_2\frac{2}{5}=0.971$$

$$-\frac{4}{4}log_2\frac{4}{4}-\frac{0}{4}log_2\frac{0}{4}=0$$

E(Outlook) =

2 *(0.971 * 5/14) = 0.694

Gain(S, Outlook) = 0.940 - 0.694 = 0.246

Chọn thuộc tính:

- Gain(S,Windy)= 0.940 – 0.892 = 0.048
- Gain(S,Outlook)=0.940 – 0.964 = 0.246

Chỉ số Gini:

- Chỉ số Gini của nút *t*:

$$GINI(t) = 1 - \sum_j p(j|t)^2$$

Trong đó $p(j|t)$ là tần suất của lớp *j* trong nút *t*

Ví Dụ:

C1	0
C2	6

p(C1) = 0/6 = 0 p(C2) = 6/6 = 1
GINI = 1 – (p(C1)²+p(C2)²) = 1 – (0+1) = 0

C1	1
C2	5

p(C1) = 1/6 p(C2) = 5/6
GINI = 1 – (1/6)² – (5/6)² = 0.278

C1	2
C2	4

p(C1) = 2/6 p(C2) = 4/6
GINI = 1 – (2/6)² – (4/6)² = 0.444

Phân nhánh bằng chỉ số GINI:

- Khi phân chia nút *p* thành *k* nhánh, chất lượng của phép chia được tính bằng:

$$GINI_{chia} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

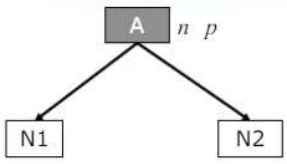
Trong đó

- n_i là số mẫu trong nút i
- n là số mẫu trong nút p

- Chọn thuộc tính có **$GINI_{chia}$ nhỏ nhất** để phân nhánh

Phân nhánh bằng thuộc tính nhị phân:

- Chỉ phân thành 2 nhánh



	p
C1	6
C2	6

Gini=0.500

	N1	N2
C1	5	1
C2	2	4

Gini=0.333

Gini(N1) =1-(5/6)²-(2/6)²
=0.194

Gini(N2) =1-(1/6)²-(4/6)²
=0.528

$Gini_{chia}$ =7/12*0.194
+5/12*0.528
=0.333

Biến đổi cây quyết định thành luật:

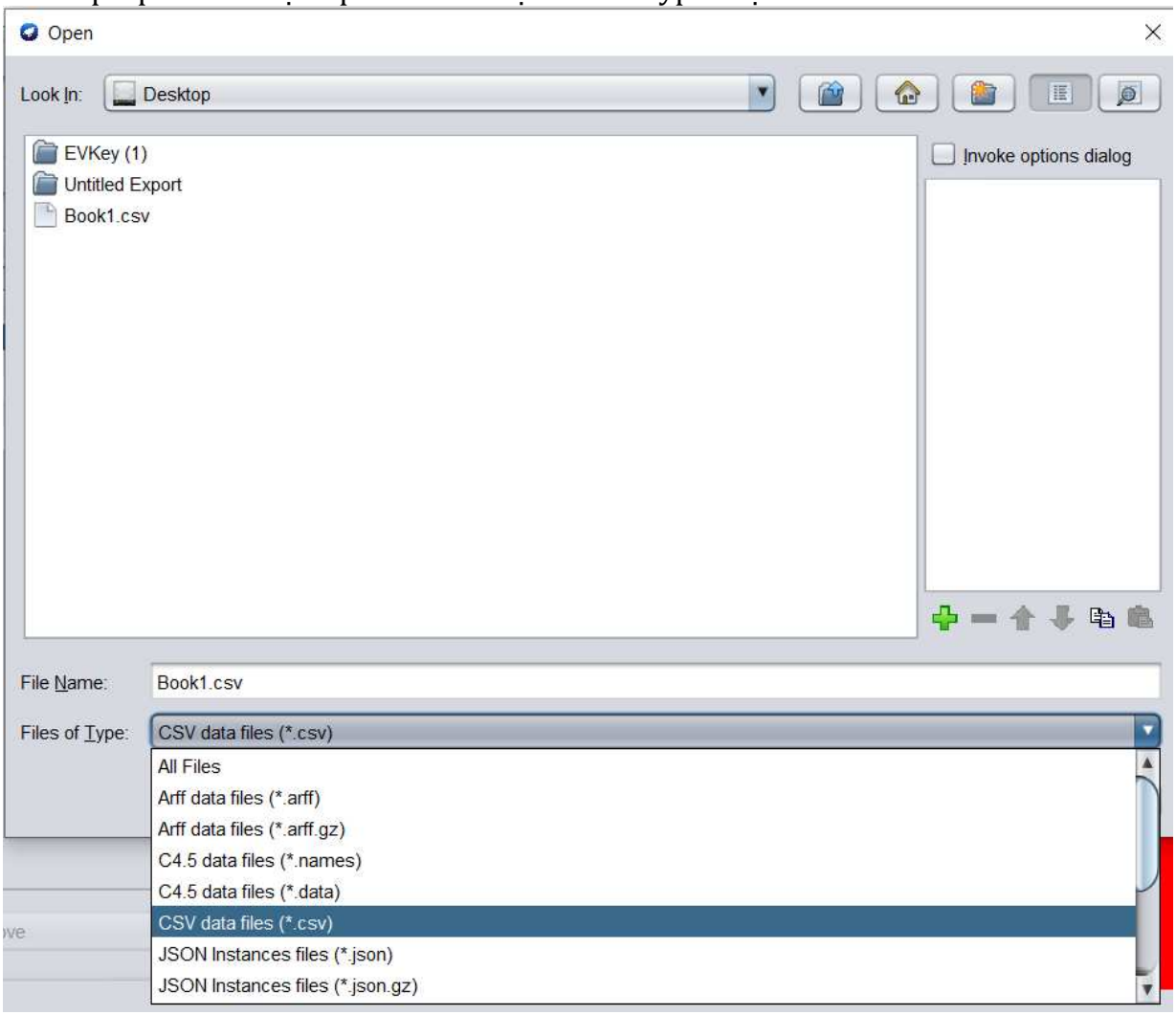
- R_1: If (Outlook = Sunny) ∧ (Humidity > 70) Then Play=No
- **R_2: If (Outlook = Sunny) ∧ (Humidity <= 70) Then Play=Yes**
- **R_3: If (Outlook = Overcast) Then Play=Yes**
- R_4: If (Outlook = Rain) ∧ (Wind = True) Then Play=No
- **R_5: If (Outlook = Rain) ∧ (Wind = False) Then Play=Yes**

Khai thác phần mềm weka và ứng dụng thuật toán ID3

-đầu tiên tạo bảng dữ liệu bằng excel như sau:

1	outlook	temperatu	humidity	windy	play
2	sunny	85	85	FALSE	dontplay
3	sunny	80	90	TRUE	dontplay
4	overcast	83	78	FALSE	play
5	rain	70	96	FALSE	play
6	rain	68	80	FALSE	play
7	rain	65	70	TRUE	dontplay
8	overcast	64	65	TRUE	play
9	sunny	72	95	FALSE	dontplay
10	sunny	69	70	FALSE	play
11	rain	75	80	FALSE	play
12	sunny	75	70	TRUE	play
13	overcast	72	90	TRUE	play
14	overcast	81	75	FALSE	play
15	rain	71	80	TRUE	dontplay

-lưu lại với định dạng csv
-mở weka , chọn explorer
-ở tab preprocess chọn open file . ở mục files of type chọn csv như sau :



-tìm đến file data csv đã tạo .
-sang tab classify chọn choose -> weka -> classicfiers-> tree -> j48
-ấn start. Cửa sổ bên phải sẽ hiện thông tin như sau :

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Book1
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
----

outlook = sunny
| humidity <= 75: play (2.0)
| humidity > 75: dontplay (3.0)
outlook = overcast: play (4.0)
outlook = rain
| windy = FALSE: play (3.0)
| windy = TRUE: dontplay (2.0)

Number of Leaves : 5

Size of the tree : 8

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	**9**	**64.2857 %**
Incorrectly Classified Instances	**5**	**35.7143 %**
Kappa statistic	**0.186**	
Mean absolute error	**0.2857**	
Root mean squared error	**0.4818**	
Relative absolute error	**60 %**	
Root relative squared error	**97.6586 %**	
Total Number of Instances	**14**	

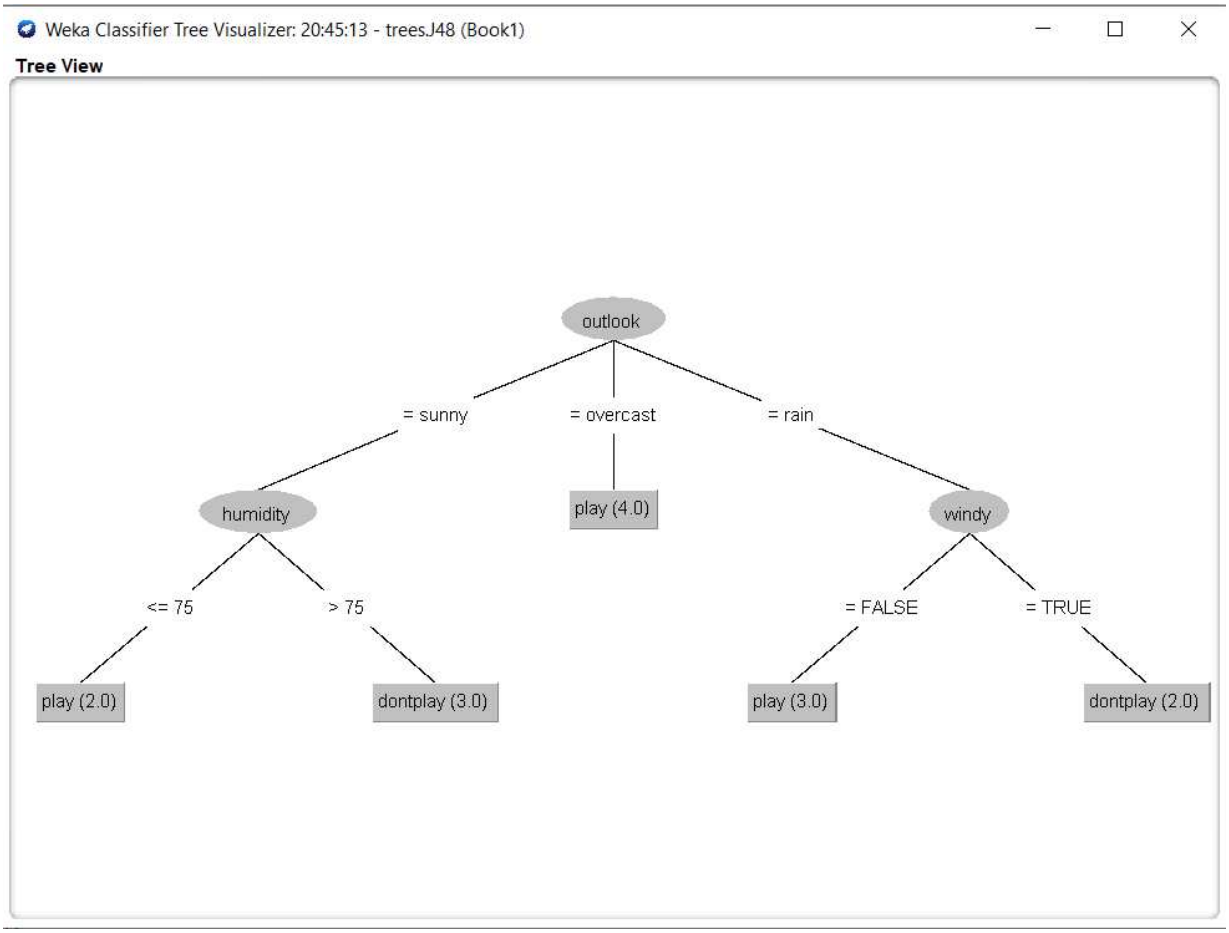
=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
	0.400	**0.222**	**0.500**	**0.400**	**0.444**	**0.189**	**0.789**	**0.738**	**dontplay**
	0.778	**0.600**	**0.700**	**0.778**	**0.737**	**0.189**	**0.789**	**0.847**	**play**
Weighted Avg.	**0.643**	**0.465**	**0.629**	**0.643**	**0.632**	**0.189**	**0.789**	**0.808**	

=== Confusion Matrix ===

a	**b**	**<-- classified as**
2	**3**	**| a = dontplay**
2	**7**	**| b = play**

- tại phần bên trái ở mục result list click chuột phải vào “trees.j48” chọn visualize trees phần mềm sẽ hiện cây quyết định :


Weka Classifier Tree Visualizer: 20:45:13 - trees.J48 (Book1)
Tree View
outlook
= sunny
= overcast
= rain
humidity
play (4.0)
windy
<= 75
> 75
= FALSE
= TRUE
play (2.0)
dontplay (3.0)
play (3.0)
dontplay (2.0)

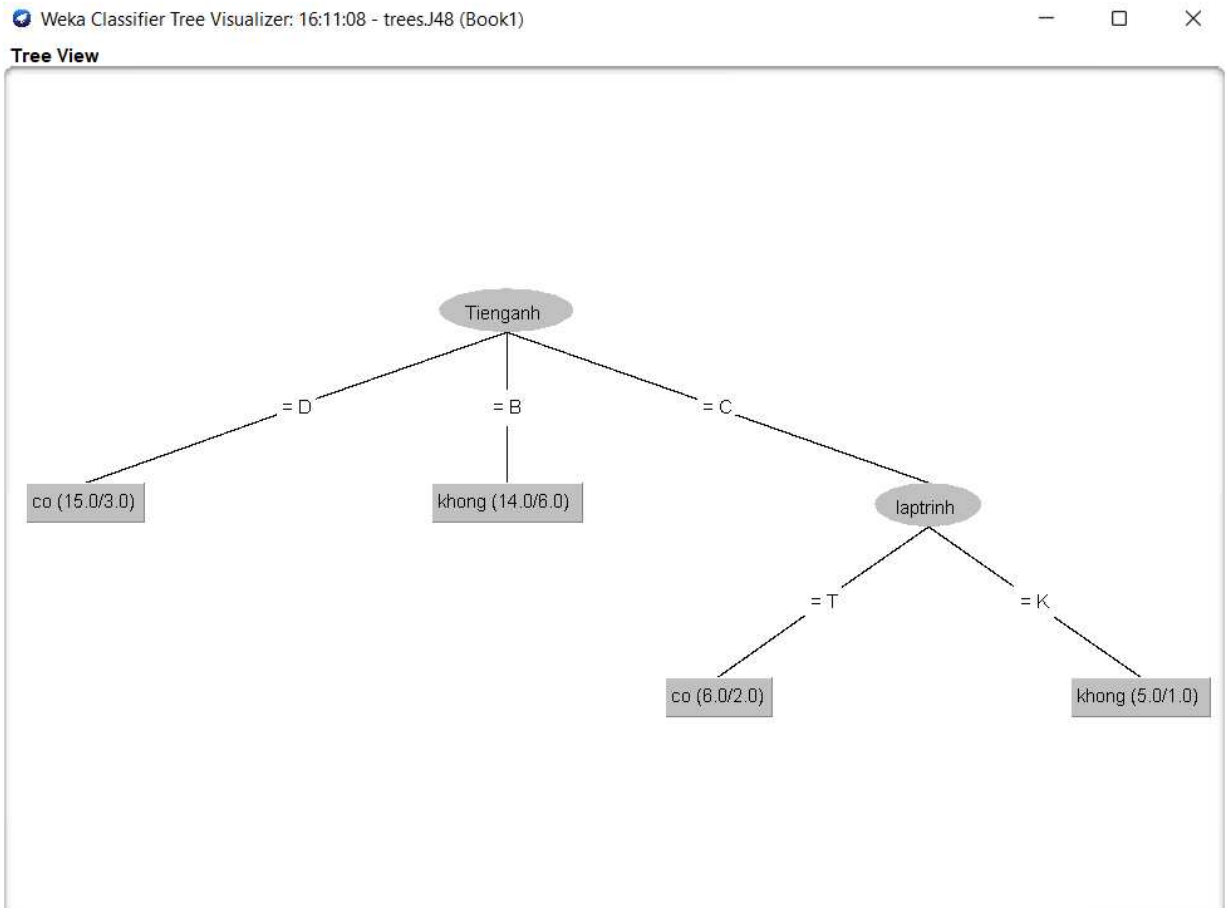
VẬN DỤNG THỰC TẾ VỀ KHẢ NĂNG CÓ VIỆC LÀM CỦA SINH VIÊN NGÀNH CNTT SAU KHI RA TRƯỜNG.

-Ta có bảng dữ liệu sau :

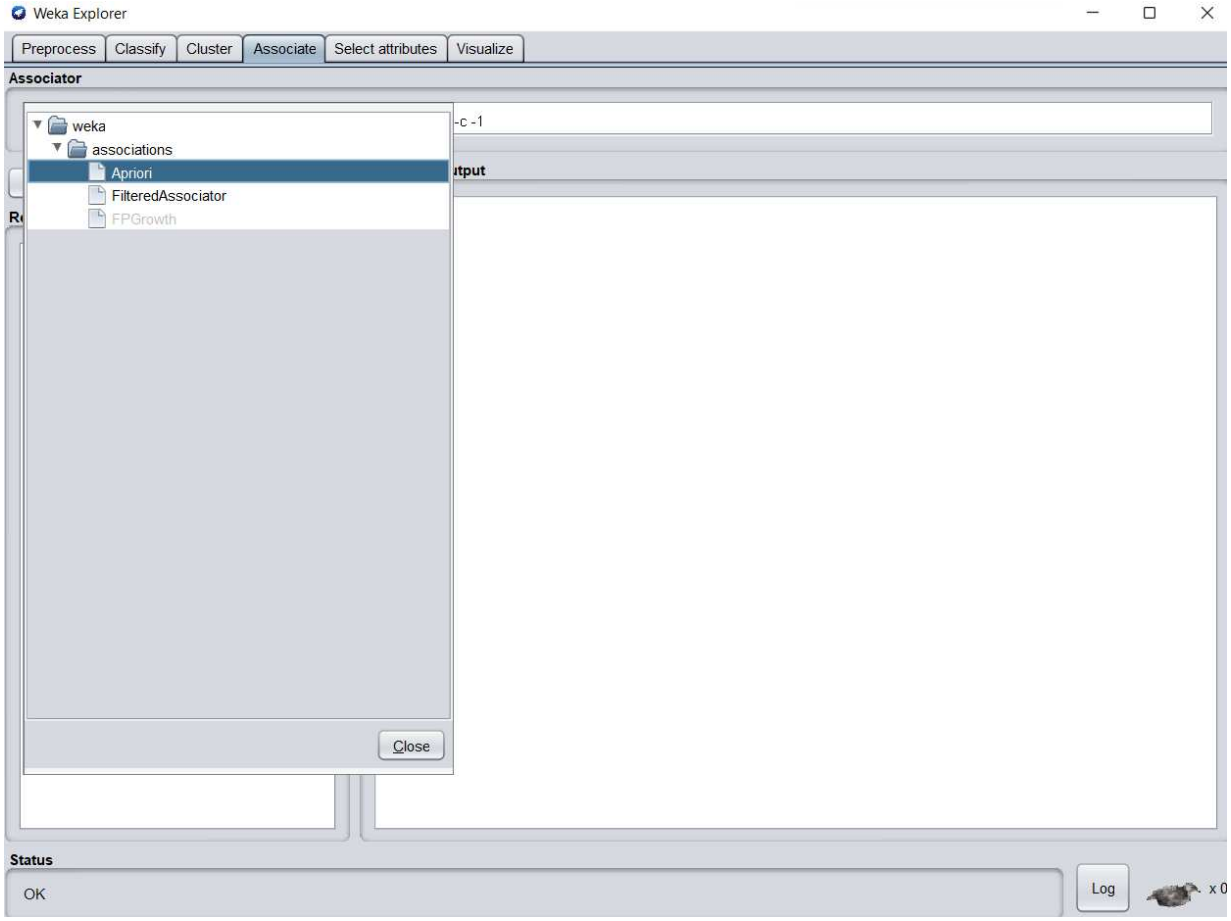
nganh	Tienganh	kynang	hocluc	laptrinh	KQ
M	D	T	TB	T	co
C	B	K	G	K	khong
C	B	T	K	T	co
K	C	K	G	T	khong
C	D	K	K	T	co
K	C	K	TB	K	khong
M	C	T	TB	T	co
M	D	T	G	K	co
K	B	T	G	K	khong
C	B	K	K	K	co
M	B	T	TB	T	co
M	C	T	G	K	khong
K	D	K	K	K	co
C	D	K	G	T	co
M	D	T	G	T	co
K	C	K	G	K	co
C	B	K	K	T	khong
M	B	K	TB	K	khong
M	B	T	K	T	khong
C	D	T	G	T	co
C	C	T	K	T	co
K	D	K	TB	K	khong
K	B	K	K	K	co
K	D	K	TB	T	khong
M	C	K	K	T	co
C	B	T	K	K	khong
M	D	K	G	K	co
M	C	T	G	K	khong
M	B	K	G	K	co
K	B	T	K	T	khong
C	C	K	K	T	khong
K	D	K	TB	K	khong
K	D	K	TB	T	co
M	D	T	K	K	co
M	B	T	K	K	co
C	C	K	TB	K	khong
C	D	T	K	T	co

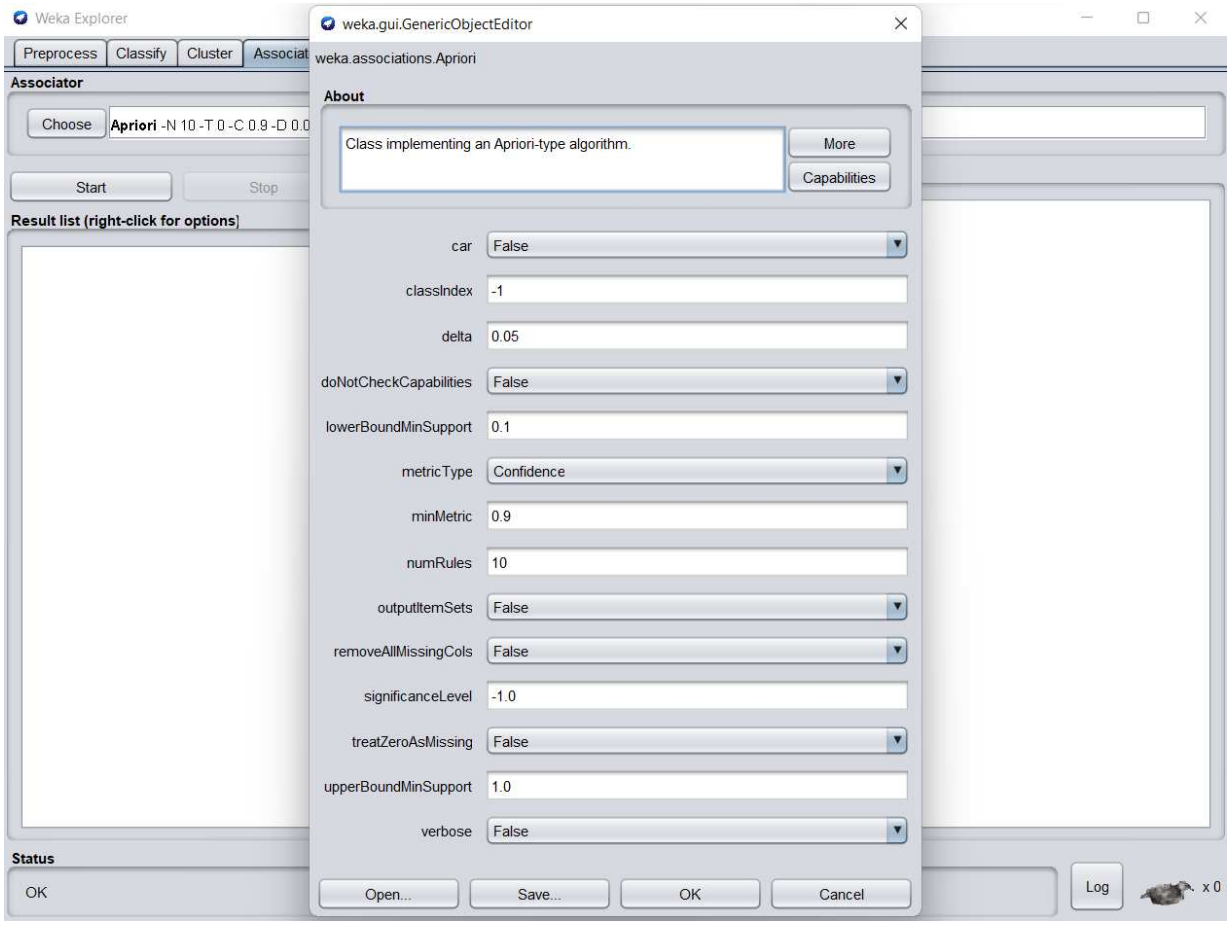
C	D	K	TB	T	co
K	C	K	K	T	co
C	B	K	K	T	khong

Tương tự như bài toán golf trên , ta cũng lưu dữ liệu dưới dạng .csv , mở trong phần mềm weka ,tuy nhiên STT không phải là thuộc tính dữ liệu nên ở tab preprocess ta đánh tick vào ô STT rồi ấn remove nó đi. Tiếp tục thực hiện các bước như trên ta được cây :

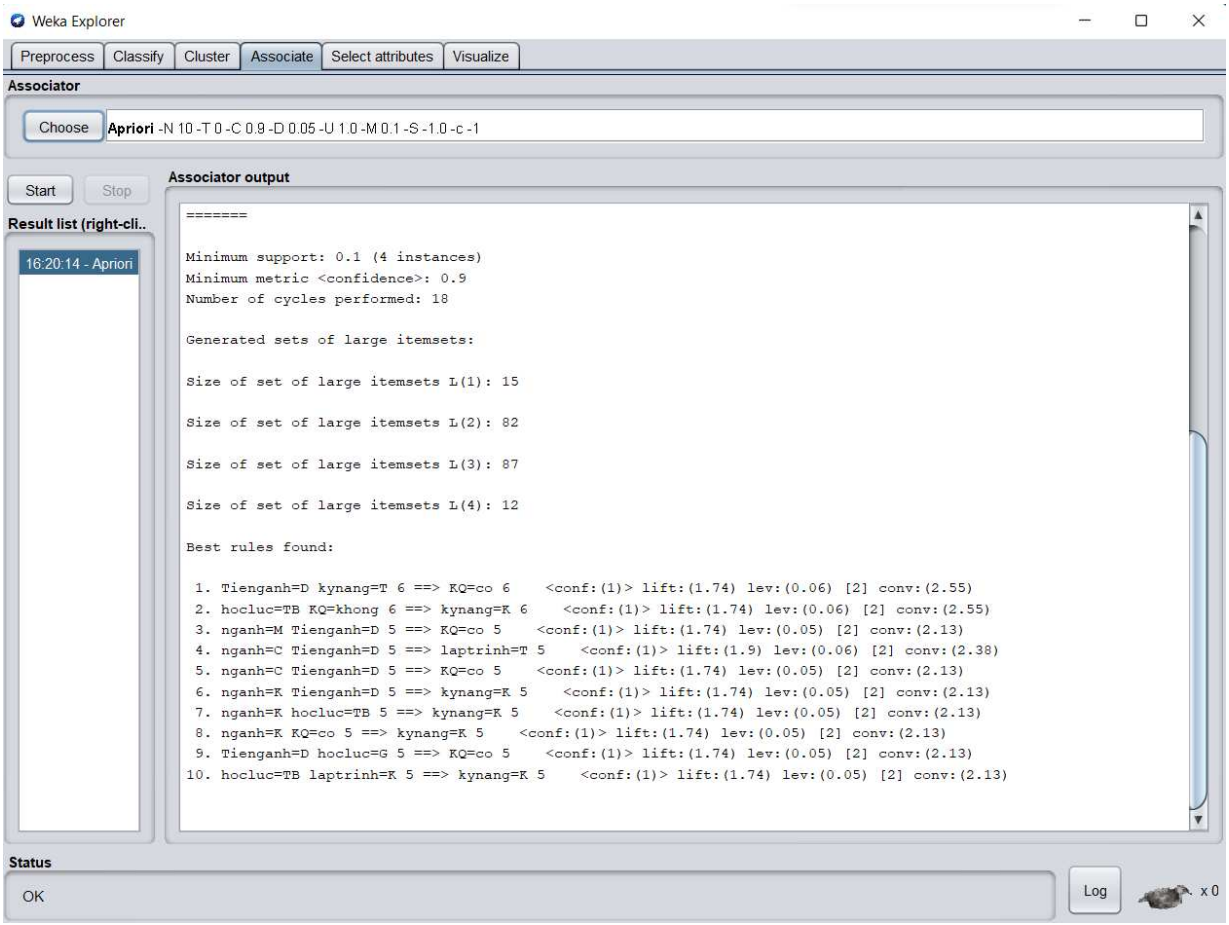


Để xây dựng tập luật , chọn tab associate, ấn choose chọn thuật toán apriori. Để nguyên các tham số mặc định như sau :


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
weka
associations
Apriori
FilteredAssociator
FPGrowth
-c -1
Close
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associat
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.0
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
More
Capabilities
car
False
classIndex
-1
delta
0.05
doNotCheckCapabilities
False
lowerBoundMinSupport
0.1
metricType
Confidence
minMetric
0.9
numRules
10
outputItemSets
False
removeAllMissingCols
False
significanceLevel
-1.0
treatZeroAsMissing
False
upperBoundMinSupport
1.0
verbose
False
Open...
Save...
OK
Cancel

Sau đó ấn start ta được , kết quả :



Như vậy với phần mềm weka ta có thể xây dựng cây quyết định với thuật toán ID3 một cách nhanh chóng .