TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

======***======

BÁO CÁO BÀI TẬP LỚN

HỌC PHẦN TRÍ TUỆ NHÂN TẠO

Đề tài: Tìm hiểu cây quyết định và ứng dụng dự đoán phá sản.

GVHD:	TS Lê Thị Thủy
Nhóm :	Nhóm 20
Thành viên:	Trần Văn Nam - 2021605962
	Hồ Mạnh Nam - 2020600924
	Mai Văn Bắc - 2021602410

Hà Nội, năm 2023

MỤC LỤC

Chương 1. Tìm hiểu Cây quyết định 3

1.1. Cây quyết định 3

Hình 1.1 là cây quyết định dự đoán phá sản..... 4

Ưu/nhược điểm của thuật toán cây quyết định 4

1.2. Tạo cây quyết định 5

1.3. Ví dụ minh hoạ..... 6

Hình 1.2. Cây quyết định sau lần phân hoạch đầu tiên 8

Hình 1.3. Cây quyết định bài toán dự đoán phá sản..... 9

Chương 2. Dự đoán phá sản 10

2.1. Bài toán 10

2.1.1. Mô tả..... 10

2.1.2. Bài toán ứng dụng dự đoán phá sản..... 10

2.2. Cơ sở dữ liệu 10

2.3. Kết quả 11

2.3.1. Chương trình:..... 11

2.3.2. Kết quả..... 13

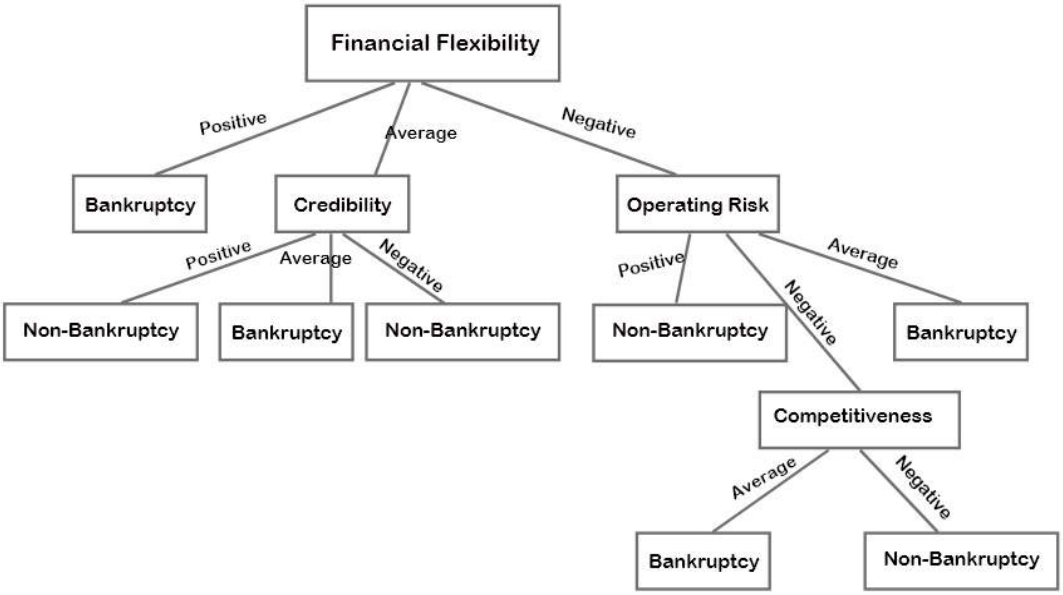
Kết luận 14

Tài liệu tham khảo.....15

Chương 1. Cây quyết định

1.1. Cây quyết định

Cây quyết định được dùng để đưa ra tập luật if – then nhằm mục đích dự báo, giúp con người nhận biết về tập dữ liệu. Cây quyết định cho phép phân loại đối tượng tùy thuộc vào các điều kiện tại các nút trong cây, bắt đầu từ gốc cây tới các nút sát lá-Nút xác định phân loại đối tượng. Mỗi nút trong của cây xác định điều kiện đối với thuộc tính mô tả của đối tượng. Mỗi nhánh tương ứng với điều kiện: Nút (thuộc tính) bằng giá trị nào đó. Đối tượng được phân loại nhờ tích hợp các điều kiện bắt đầu từ nút gốc của cây và các thuộc tính mô tả với giá trị của thuộc tính đối tượng.



Hình 1.1 là cây quyết định dự đoán phá sản

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. Thậm 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.

1.2. Tạo cây quyết định

Xét bảng dữ liệu T = (A, D) trong đó A = {A_1, A_2,..., A_n } là tập thuộc tính dẫn xuất, D = {r_1, r_2, ..., r_n} là thuộc tính mục tiêu. Vấn đề đặt ra là trong tập thuộc tính A ta phải chọn thuộc tính nào để phân hoạch? Một trong các phương pháp đó là dựa vào độ lợi thông tin. Hay còn gọi là thuật giải ID3.

Lựa chọn chủ yếu trong giải thuật ID3 là chọn thuộc tính nào để đưa vào mỗi nút trong cây. Ta sẽ chọn thuộc tính phân rã tập mẫu tốt nhất. Thước đo độ tốt của việc chọn lựa thuộc tính là gì? Ta cần xác định một độ đo thống kê, gọi là thông tin thu được, đánh giá từng thuộc tính được chọn tốt như thế nào còn phụ thuộc vào việc phân loại mục tiêu của tập mẫu. ID3 sử dụng thông tin thu được đánh giá để chọn ra thuộc tính cho mỗi bước giữa những thuộc tính ứng viên, trong quá trình phát triển cây.

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Để đánh giá chính xác thông tin thu được, dùng Entropy(S): Độ bất định (độ pha trộn/độ hỗn tạp) của S liên quan đến sự phân loại đang xét

Trong đó p_i là xác suất xuất hiện trạng thái i của hệ thống. Theo lý thuyết thông tin: mã có độ dài tối ưu là mã gán $-log_2 p$ bits cho thông điệp có xác suất là p. S là một tập huấn luyện.

Nếu gọi p là xác suất xuất hiện các ví dụ dương trong tập S, p là xác suất xuất hiện các ví dụ âm trong tập S. Entropy đo độ bất định của tập S sẽ là:

Entropy (S) $= -p \ log_2 p \ - p \ log_2 p$

Quy định 0.log 0 = 0

Chẳng hạn với tập S gồm 14 mẫu có chung một vài giá trị logic gồm 9 mẫu dương và 5 mẫu âm. Khi đó đại lượng Entropy của tập S liên quan đến sự phân loại logic này là:

$$\text{Entropy}([9+, 5-]) = -(9/14)\, log_2\,(9/14) - (5/14)\, log_2\,(5/14) = 0{,}940$$

Chú ý :

Đại lượng Entropy = 0 nếu tất cả thành viên của tập S cùng thuộc một lớp (vì nếu tất cả là dương (P+ = 1), do đó P- = 0, Entropy(S) = $-1log_2 1 - 0log_2 0 = 0$).

Đại lượng Entropy(S) = 1 khi tập S chứa tỉ lệ tập mẫu âm và mẫu dương là như nhau. Nếu tập S chứa tập mẫu âm và tập mẫu dương có tỉ lệ P+ khác P- thì Entropy(S) ∈ (0,1).

Dựa trên sự xác định entropy, ta tính Gain(S, A) = Lượng giảm entropy mong đợi qua việc chia các ví dụ theo thuộc tính A

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{values(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

1.3. Ví dụ minh hoạ

Xem xét nhiệm vụ học được đưa ra bởi tập mẫu dưới đây , thuộc tính mục tiêu ở đây là: class có giá trị là N hoặc NB, giá trị thuộc tính này dự đoán dựa vào các thuộc tính mô tả.

Kí hiệu: P=Positive,A-Average,N-negative,B-Bankruptcy,NB-Non-Bankruptcy

Financial_Flexibility : khả năng linh hoạt tài chính

Credibility : Sự uy tín

Competitiveness : khả năng cạnh tranh

Operating_Risk : Rủi ro hoạt động

Bảng tập mẫu:

Object	Financial_Flexibility	Credibility	Competitiveness	Operating_Risk	Class
D1	N	P	P	P	NB
D2	P	P	P	N	B
D3	P	P	P	N	B

D4	A	P	P	P	NB
D5	P	A	N	P	B
D6	A	N	P	N	NB
D7	A	P	N	P	NB
D8	N	N	A	A	B
D9	N	N	A	N	B
D10	N	A	N	N	NB
D11	A	A	P	N	B
D12	N	N	N	N	NB
D13	P	A	N	A	B
D14	N	N	A	P	NB

Bước 1: Tạo nút đỉnh cho cây quyết định:

Giá trị thông tin thu được cho mỗi thuộc tính:

Entropy(S)=1

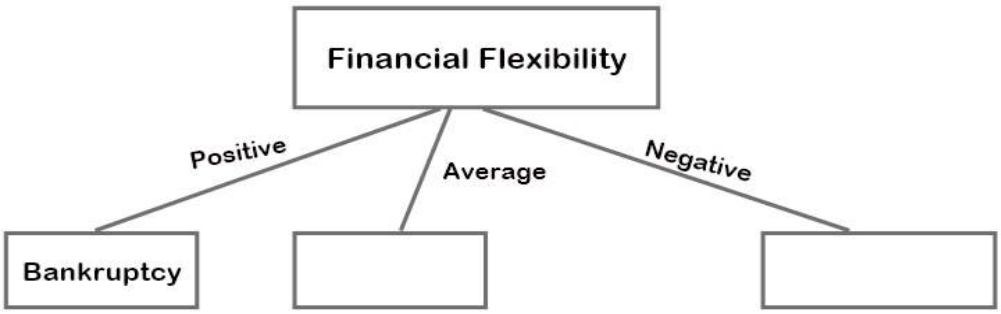
Gain(S, Financial Flexibility)=0.401

Gain(S, Credibility)=0.044

Gain(S, Competitiveness)=0.028

Gain(S, Operating Risk)=0.25

Theo đánh giá thông tin thu được, thuộc tính Financial Flexibility cung cấp dự đoán tốt nhất về thuộc tính mục tiêu “Class” trên tập mẫu. Do đó, thuộc tính “Financial Flexibility” được chọn là thuộc tính quyết định cho nút gốc, nhánh được tạo ra dưới nút gốc tương ứng với mỗi giá trị của thuộc tính như Credibility, Competitiveness, Operating Risk cùng với tập mẫu sẽ thêm vào mỗi nút con mới.



Hình 1.2. Cây quyết định sau lần phân hoạch đầu tiên

Bước 2: Xét nhánh Average với các thuộc tính còn lại trong bảng.

Entropy($S_{Average}$)= 0.811

Gain ($S_{Average,}$ Credibility)=0.811

Gain($S_{Average,}$ Competitiveness)=0.123

Gain($S_{Average,}$ Operating Risk)=0.311

Theo đánh giá thông tin thu được, thuộc tính Credibility cung cấp dự đoán tốt nhất.

Chọn Credibility là nút tiếp sau Average.

Bước 3: Xét nhánh Negative với các thuộc tính còn lại trong bảng.

Entropy($S_{Negative}$)= 0.918

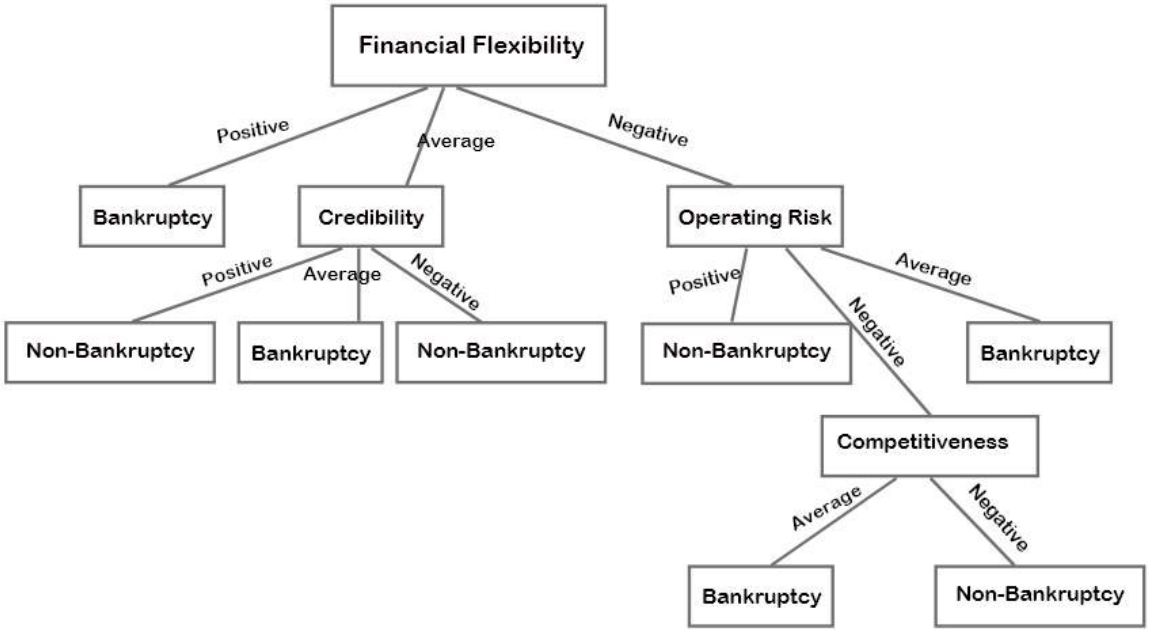
Gain ($S_{Negative,}$ Credibility)=0.251

Gain($S_{Negative,}$ Competitiveness)=0.251

Gain($S_{Negative,}$ Operating Risk)=0.459

Theo đánh giá thông tin thu được, thuộc tính Operating Risk cung cấp dự đoán tốt nhất.

Chọn Operating Risk là nút tiếp sau Negative.



Hình 1.3. Cây quyết định bài toán dự đoán phá sản

Chương 2. Dự đoán phá sản

2.1. Bài toán

2.1.1. Mô tả

- Sử dụng ID3 (Iterative Dichotomiser 3) → dùng Entropy function và Information gain

-Lựa chọn chủ yếu trong giải thuật ID3 là chọn thuộc tính nào để đưa vào mỗi nút trong cây. Ta sẽ chọn thuộc tính phân rã tập mẫu tốt nhất.

-ID3 sử dụng thông tin thu được đánh giá để chọn ra thuộc tính cho mỗi bước giữa những thuộc tính ứng viên, trong quá trình phát triển cây.

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

*pi là xác suất xuất hiện trạng thái i của hệ thống

-Để đánh giá chính xác thông tin thu được, dùng Entropy(S): Độ bất định (độ pha trộn/độ hỗn tạp) của S liên quan đến sự phân loại đang xét

-Dựa trên sự xác định entropy, ta tính Gain(S, A) = Lượng giảm entropy mong đợi qua việc chia các ví dụ theo thuộc tính A

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

2.1.2. Bài toán ứng dụng dự đoán phá sản

-Dữ liệu bài toán T = (A,D)

+A là tập các thuộc tính dẫn xuất. A(Financial Flexibility, Credibility, Competitiveness, Operating Risk)

+D là thuộc tính mục tiêu D = Class

2.2. Cơ sở dữ liệu

-Nguồn:https://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy

-Data set gồm 4 thuộc tính dẫn xuất (đã bỏ 2 thuộc tính) và 1 thuộc tính mục tiêu

+ Thuộc tính dẫn xuất:Financial Flexibility (tính linh hoạt tài chính), Credibility (Sự uy tín), Competitiveness(Khả năng cạnh tranh), Operating risk(rủi ro hoạt động). Thuộc tính có các giá trị P(Positive : Tốt), A(Average: Trung bình) và N(Negative:Kém)

  +Thuộc tính đích: class gồm 2 giá trị B(Bankrupt) và NB(Non-Bankrupt)

2.3. Kết quả

2.3.1. Chương trình:

B1, Import thư viện cần thiết

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
```

- Thư viện pandas để đọc dữ liệu
- DecisionTreeClassifier của thư viện sklearn dùng để giải quyết bài toán cây quyết định

B2, Đọc data set

```
dataset = pd.read_csv("Qualitative_Bankruptcy.data.txt",
names = ["Industrial_Risk","Management_Risk","Financial_Flexibility","Credibility","Competitiveness","Operating_Risk","Class"])
```

B3, Xử lý dữ liệu trước khi thực hiện bài toán

- Drop 2 thuộc tính giúp dataset trực quan, dễ hiểu và dễ hình dung

  +Method drop của pandas.dataFrame với tham số axis có 2 giá trị 0 hay ‘index’ và 1 hay ‘columns’

```
dataset = dataset.drop("Industrial_Risk",axis = 1)
dataset = dataset.drop("Management_Risk",axis = 1)
```

- Convert dữ liệu kiểu chuỗi về kiểu số

  +Method map của pandas.dataFrame lọc từng giá trị của column và convert theo tham số truyền vào

```
Financial_Flexibility= {'P':0,'A':1,'N':2}
Credibility= {'P':0,'A':1,'N':2}
Competitiveness= {'P':0,'A':1,'N':2}
Operating_Risk= {'P':0,'A':1,'N':2}
Class = {'B':0,'NB':1}

dataset['Financial_Flexibility']                                    =
dataset['Financial_Flexibility'].map(Financial_Flexibility)
dataset['Credibility'] = dataset['Credibility'].map(Credibility)
dataset['Competitiveness'] = dataset['Competitiveness'].map(Competitiveness)
dataset['Operating_Risk'] = dataset['Operating_Risk'].map(Operating_Risk)
dataset['Class'] = dataset['Class'].map(Class)
```

B4, Chia dữ liệu để train và test

-Hàm iloc dùng để chia dataFrame.

+[:20,:-1] lấy 20 dòng đầu và bỏ cột cuối

+[20:,:-1] bỏ 20 dòng đầu và lấy toàn bộ số dòng còn lại, bỏ cột cuối

+[:20,-1] lấy 20 dòng đầu và chỉ lấy cột cuối

+[20:,-1] bỏ 20 dòng đầu và lấy toàn bộ số dòng còn lại, chỉ lấy cột cuối

```
train_features = dataset.iloc[:20,:-1]
test_features = dataset.iloc[20:,:-1]
train_targets = dataset.iloc[:20,-1]
test_targets = dataset.iloc[20:,-1]
```

B5, Train dữ liệu

-Hàm DecisionTreeClassifier dùng để giải bài toán cây quyết định

+Tham số criterion là tiêu chí đo lường của một lần tách. Criterion có 2 giá trị gini(default) và entropy

+Entropy: là một thước đo thông tin chỉ ra sự rối loạn của các thuộc tính dẫn xuất với mục tiêu,

+Information entropy : thể hiện mức độ hỗn loạn hay độ nhiễu của data dùng để xác định thuộc tính nào mang lại nhiều thông tin hơn

-Ta dùng method fit để train dữ liệu với train features là dữ liệu về các thuộc tính dẫn xuất, train_targers là dữ liệu thuộc tính mục tiêu tương ứng

```
tree = DecisionTreeClassifier(criterion = 'entropy').fit(train_features,train_targets)
```

B6, Dự đoán và in ra

-Sau khi train dữ liệu ta có tree, dùng method predict của tree để dự đoán dữ liệu thuộc tính mục tiêu tương ứng(prediction) với dữ liệu của các thuộc tính dẫn xuất được truyền vào(test_features)

```
prediction = tree.predict(test_features)
print(prediction)
```

B7, In ra độ chính xác của chương trình

-Dùng method score để tính độ chính xác của dự đoán bằng cách truyền 2 tham số

+Tham số thứ nhất : dữ liệu của các thuộc tính dẫn xuất(test_features)

+Tham số thứ hai : dữ liệu của thuộc tính mục tiêu

```
print("The prediction accuracy is:",tree.score(test_features,test_targets)*100,"%")
```

2.3.2. Kết quả

```
:\My Drive\Study\AI\BTL> & "C:/Program Files/Python39/python.exe" "g:/My Drive/Study/AI/BTL/test.py"
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 1]
prediction accuracy is:  98.26086956521739 %
:\My Drive\Study\AI\BTL>
```

Kết luận

Trong quá trình thực hiện đồ án tốt nghiệp, nhóm đã cố gắng hết sức để tìm hiểu và học hỏi nhưng vì khả năng còn giới hạn không tránh khỏi những sai sót, nên có thể chưa giải quyết được tất cả những vấn đề đặt ra. Rất mong nhận được sự thông cảm của quý thầy cô và các bạn. Em xin chân thành cảm ơn.

Những kết quả đạt được:

- Sự hiểu biết về thuật toán Decision Tree cơ bản tương đối tốt
- Từ những gì đã làm được, từ đó biết thêm về AI, ứng dụng của AI vào cuộc sống công nghệ hiện đại
- Làm quen ngôn ngữ lập trình Python

Tài liệu tham khảo

https://python-course.eu/Decision_Trees.php

Dataset : https://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy