ISSN 2354-0575

PHÂN LỚP HỌC SINH VÀ PHÂN LOẠI CÁN BỘ SỬ DỤNG CÁC THUẬT TOÁN TRONG HỌC MÁY

Nguyễn Quang Hoan[1], Lê Thị Tuyết Mây[2], Nguyễn Mạnh Tuân[3], Nguyễn Ngọc Ánh[4]
1 Trường Đại học Sư phạm Kỹ thuật Hưng Yên
2 Trường Cao đẳng Nghề Điện Biên
3 Trung tâm Ngoại ngữ - Tin học tỉnh Điện Biên
4 Trường Cao đẳng Kinh tế - Kỹ thuật Điện Biên

Ngày tòa soạn nhận được bài báo: 17/09/2017
Ngày phản biện đánh giá và sửa chữa: 10/11/2017
Ngày bài báo được chấp nhận đăng: 25/11/2017

Tóm tắt:
Bài báo tiến hành phân tích, xử lý dữ liệu, chọn lựa thuật toán ID3, C4.5, Bayes ứng dụng phân lớp học viên của Trung tâm Ngoại ngữ - Tin học, phân loại cán bộ Trường Cao đẳng Nghề Điện Biên và thử nghiệm trên phần mềm Weka. Bộ tiêu chí đánh giá chất lượng phân lớp cho cán bộ trường, cho việc phân chia được thử nghiệm, đánh giá.
Từ khóa: *Thuật toán ID3, thuật toán Bayes, Độ lợi thông tin.*

1. Giới thiệu

Có nhiều thuật toán phân lớp như: Cây quyết định (Thuật toán Quinlan, ID3, Độ lộn xộn, C4.5, C5.0…; K-NN (K-Nearest Neighbor); Bayes; Mạng nơron; Hệ mờ… Mỗi thuật toán có ưu điểm, hạn chế, độ phức tạp, đối tượng ứng dụng khác nhau [10]. Cây quyết định: Đơn giản, nhanh, hiệu quả và được ứng dụng thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại văn bản [2], [9]… Thuật toán Bayes cho kết quả tốt trong thực tế, mặc dù chịu những giả thiết về tính độc lập xác suất của các thuộc tính và được ứng dụng trong các bài toán dự đoán, phân loại văn bản, Spam…[5], [8]. Trong bài báo này, chúng tôi sử dụng thuật toán C4.5 và Bayes để chia lớp cho học viên của Trung tâm Ngoại ngữ - Tin học và Trường Cao đẳng Nghề.

2. Các thuật toán cho bài phân lớp

Từ các phân tích trên, với quy mô dữ liệu không lớn, độ chính xác không đòi hỏi cao đối với một trường nghề và trung tâm; có thể dùng thuật toán ID3 ([1], [7]) C4.5 (J48) và Bayes ([5], [6]) để cho phân loại chất lượng học sinh hoặc chia lớp cho học viên.

2.1. Thuật toán ID3

2.1.1. Thuật toán ID3

Đầu vào: Tập dữ liệu huấn luyện gồm các thuộc tính tình huống, hay đối tượng nào đó, và giá trị dùng để phân loại.

Đầu ra: Cây quyết định có khả năng phân loại đúng đắn các ví dụ trong tập dữ liệu huấn luyện, và có thể là phân loại đúng cho cả các ví dụ không có trong tập huấn luyện hay chưa gặp trong tương lai.

Bắt đầu với nút gốc:

Bước 1: *Chọn thuộc tính quyết định "tốt nhất" cho nút gốc; gán nó cho A.*

Bước 2: *Với mỗi giá trị của A, tạo nhánh*

Bước 3: *Lặp lại Bước 1 và 2 cho nhánh*

Bước 4: *Nếu các mẫu huấn luyện trong nhánh được phân loại đồng nhất: DỪNG, được nút lá. Ngược lại, lặp từ 1 cho đến 4.*

- Công thức làm tiêu chí quyết định

+ Entropy của một tập S có 2 phân lớp

$$Entropy(S) = - P_+ log_2 P_+ - P_- log_2 P_- \quad (2.1)$$

+ Entropy của tập S có *c* phân lớp

$$Entropy(S) = \sum_{i=1}^{c} -P_i log_2 P_i \quad (2.2)$$

- Tiêu chí quyết định: độ lợi lớn nhất

Tập dữ liệu *S* gồm có *n* thuộc tính A_i ($i = 1, 2,..., n$) giá trị Độ lợi thông tin ($Gain(S, A)$) của *A*

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2.3)$$

2.1.2. Thử nghiệm bài toán chia lớp Ngoại ngữ

a. Phân tích bài toán

Trung tâm Ngoại ngữ - Tin học tỉnh Điện Biên hàng năm tuyển sinh (số lượng như Bảng 1) với 4 đặc trưng chính ảnh hưởng đến chia lớp cho học viên đăng ký học Ngoại ngữ: Trình độ chuyên môn ***(TDCM)***, Cấp trường ***(CTr)***, Chức danh nghề nghiệp giáo viên ***(Hang)*** và Đăng ký học tiếng Anh

theo cấp độ ***(DK)***. Mỗi đặc trưng có những giá trị khác nhau. Từ đó ta xây dựng bài toán, chia dữ liệu đầu vào thành 4 đặc trưng:

Bảng 1. *Bảng cơ sở phân lớp Ngoại ngữ*

TDCM	CTr	Hang	QD
TC	MN	IV	A1
CD	TH	III	A2
DH	THCS	II	B1
ThS	THPT	I	
Số lượng (SL): 4	***SL: 4***	***SL: 4***	***SL: 3***

+ ***TDCM***: Là trình độ đào tạo của các cán bộ, viên chức gồm 4 loại: ThS (Thạc sĩ), DH (Đại học), CD (Cao đẳng), TC (Trung cấp).

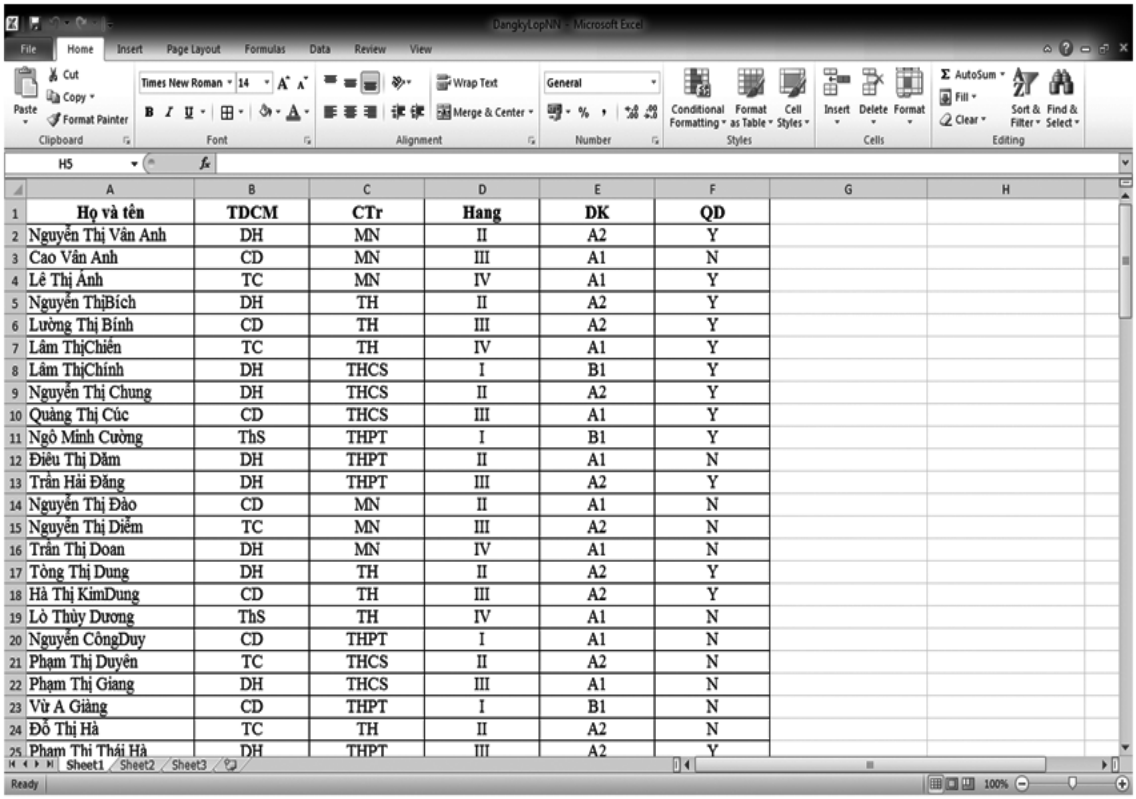
+ ***CTr***: Gồm 4 loại: MN (Cấp Mầm non), TH (Cấp Tiểu học), THCS (Cấp Trung học cơ sở), THPT (Cấp Trung học phổ thông).

+ ***Hang***: Gồm 4 loại: I (Hạng I), II (Hạng II), III (Hạng III), IV (Hạng IV).

+ ***DK***: Gồm 3 loại A1, A2, B1 (Trình độ tiếng Anh theo khung tham chiếu Châu Âu).

b. Thử nghiệm bài toán

Sau khi phân tích dữ liệu và tìm hiểu thuật toán, chúng tôi tiến hành thử nghiệm bài toán trên phần mềm *Weka*, một phần mềm khá nổi tiếng cho khai phá dữ liệu.

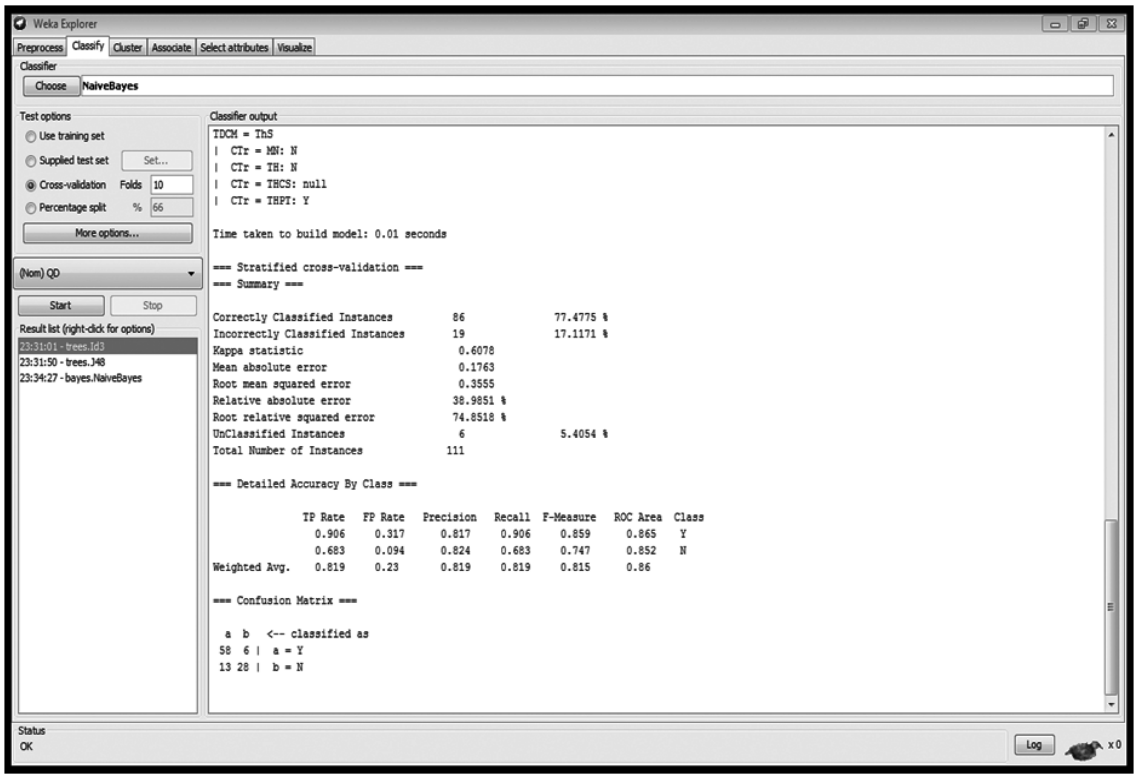


	Họ và tên	TDCM	CTr	Hang	DK	QD
2	Nguyễn Thị Vân Anh	DH	MN	II	A2	Y
3	Cao Văn Anh	CD	MN	III	A1	N
4	Lê Thị Ánh	TC	MN	IV	A1	Y
5	Nguyễn ThịBích	DH	TH	II	A2	Y
6	Lường Thị Bình	CD	TH	III	A2	Y
7	Lâm ThịChiến	TC	TH	IV	A1	Y
8	Lâm ThịChính	DH	THCS	I	B1	Y
9	Nguyễn Thị Chung	DH	THCS	II	A2	Y
10	Quàng Thị Cúc	CD	THCS	III	A1	Y
11	Ngô Minh Cường	ThS	THPT	I	B1	Y
12	Điêu Thị Dăm	DH	THPT	II	A1	N
13	Trần Hải Đăng	DH	THPT	III	A2	Y
14	Nguyễn Thị Đào	CD	MN	II	A1	N
15	Nguyễn Thị Diễm	TC	MN	III	A2	N
16	Trần Thị Doan	DH	MN	IV	A1	N
17	Tòng Thị Dung	DH	TH	II	A2	Y
18	Hà Thị KimDung	CD	TH	III	A2	Y
19	Lò Thùy Dương	ThS	TH	IV	A1	N
20	Nguyễn CôngDuy	CD	THPT	I	A1	N
21	Phạm Thị Duyên	TC	THCS	II	A2	N
22	Phạm Thị Giang	DH	THCS	III	A1	N
23	Vừ A Giàng	CD	THPT	I	B1	N
24	Đỗ Thị Hà	TC	TH	II	A2	N
25	Phạm Thị Thái Hà	DH	THPT	III	A2	Y

Hình 1. *Bảng dữ liệu hosodangkyNNTH.csv*

Sau đó, ta tiến hành tiền xử lý dữ liệu với phần mềm *Weka* để lựa chọn các thuộc tính cần thiết và loại bỏ các thuộc tính không cần thiết để phục vụ cho quá trình phân loại.

Sau khi thử nghiệm bài toán trên phần mềm Weka sử dụng thuật toán ID3 chúng ta được kết quả như sau (Hình 2):



Hình 2. *Kết quả dự đoán theo Weka*

Kết quả: Sau khi sử dụng thuật toán ID3 ta rút ra kết quả từ 111 bản ghi trong tập dữ liệu *ToanTruong.CSV*.

2.2. Giải thuật cây quyết định C4.5 (J48)

2.2.1. Thuật toán C4.5 (J48)

C4.5 là thuật toán cải tiến từ ID3 nên các bước tương tự, chỉ khác về tiêu chí, công thức chọn nút gốc. C4.5 sử dụng cơ chế lưu trữ dữ liệu thường trú trong bộ nhớ, chính đặc điểm này làm C4.5 chỉ thích hợp với những cơ sở dữ liệu nhỏ, và cơ chế sắp xếp lại dữ liệu tại mỗi node trong quá trình phát triển cây quyết định. C4.5 còn chứa một kỹ thuật cho phép biểu diễn lại cây dưới dạng danh sách sắp thứ tự các luật **if-then**.

Để đánh giá và chọn thuộc tính khi phân hoạch dữ liệu, Quinlan đề nghị sử dụng độ lợi thông tin (chọn thuộc tính có độ lợi thông tin lớn nhất) và tỉ số độ lợi dựa trên hàm entropy của Shannon. Độ lợi thông tin của một thuộc tính được tính bằng: độ đo hỗn loạn trước khi phân hoạch trừ cho sau khi phân hoạch. Giả sử P_i là xác xuất mà phần tử trong dữ liệu S thuộc lớp C_i ($i = 1, k$), đo độ hỗn loạn thông tin trước khi phân hoạch được tính theo công thức (2.4):

$$I(S) = -\sum_{i=1}^{k} P_i \log_2(P_i) \quad (2.4)$$

Độ đo hỗn loạn sau khi sử dụng thuộc tính A phân hoạch dữ liệu S thành v phần được tính như công thức (2.5):

$$I_A(S) = \sum_{i=1}^{v} \frac{|S_i|}{|S|} \times I(S_i) \quad (2.5)$$

Độ lợi thông tin khi chọn thuộc tính A phân hoạch dữ liệu S thành v phần được tính theo công thức (2.6) [2]:

$$G(S) = I(S) - I_A(S) \quad (2.6)$$

Tuy nhiên, khi dữ liệu có thuộc tính có nhiều giá trị hơn các thuộc tính khác, độ lợi thông tin tăng trên các thuộc tính có nhiều giá trị phân hoạch. Để giảm bớt sự lệch này, Quinlan cũng đề nghị sử dụng tỉ số độ lợi. Tỉ số độ lợi tính đến số lượng và độ lớn của các nhánh khi chọn một thuộc tính phân hoạch, được tính bằng độ lợi thông tin chia cho thông tin của phân phối dữ liệu trên các nhánh. Giả sử khi sử dụng thuộc tính A phân hoạch dữ liệu S thành v phần, thông tin của phân phối dữ liệu được tính:

$$P(S) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{S} \quad (2.7)$$

Tỉ số độ lợi được tính như công thức (2.8):

$$GainRatio(S) = \frac{G(S)}{P(S)} \quad (2.8)$$

Trong mô hình phân lớp C4.5, có thể dùng một trong hai loại chỉ số *Information Gain* hay *Gain ratio* (mặc định) để xác định thuộc tính tốt nhất.

2.2.2. Xử lý những giá trị thiếu trong C4.5

Dữ liệu thiếu là giá trị của thuộc tính không xuất hiện trong một vài trường hợp có thể do lỗi trong quá trình nhập bản ghi vào cơ sở dữ liệu hoặc giá trị của thuộc tính đó được đánh giá là không cần thiết trong những trường hợp đó. Trong quá trình xây dựng cây từ tập dữ liệu đào tạo *S, B* là **test** dựa trên thuộc tính A_a với các giá trị đầu ra là $b_1,b_2,...,b_t$. Tập S_0 là tập con các trường hợp trong *S* mà có giá trị thuộc tính A_a không biết và S_i biểu diễn các trường hợp với đầu ra là b_i trong test *B*. Khi đó độ đo độ lợi thông tin của test *B* giảm vì chúng ta không phân được lớp nào từ các trường hợp trong S_0 và được tính theo:

$$G(S,B) = \frac{|S-S_0|}{S} G(S-S_0,B) \quad (2.9)$$

Trong đó:

S là tập dữ liệu huấn luyện

B là tập dữ liệu test

Tập con S_0 là tập con các trường hợp trong *S* có giá trị thuộc tính A_a không biết

S_i biểu diễn các trường hợp với đầu ra b_i trong *B*

Từ đó *P*(*S, B*) cũng thay đổi như sau:

$$P(S,B) = -\frac{|S_0|}{S}\log_2\left(\frac{|S_0|}{S}\right) - \sum_{i=1}^{t}\frac{|S_i|}{S}\log_2\left(\frac{|S_i|}{S}\right) \quad (2.10)$$

Hai thay đổi này làm giảm giá trị của **Test** liên quan đến thuộc tính có tỉ lệ giá trị thiếu cao. Nếu **Test***B* được chọn, C4.5 không tạo nhánh riêng trên cây quyết định cho S_0. Thay vào đó, thuật toán có cơ chế phân chia các trường hợp trong S_0 về các tập con S_i là tập con mà có giá trị thuộc tính test xác định theo trọng số:

$$\frac{|S_i|}{|S-S_0|}$$

2.3. Thuật toán Bayes [5], [6]

Giả sử *D* là tập huấn luyện gồm các mẫu $X=(x_1, x_2,...,x_n)$. $C_{i,D}$ là tập các mẫu của *D* thuộc lớp C_i $(i=\{1,...,m\})$.

Các thuộc tính $(x_1, x_2,...,x_n)$ với giả thiết độc lập nhau được tính như sau:

$$P = (C_i \mid X) = \prod_{i=1}^{n} P(x_k \mid C_i)$$
$$= P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times ... \times P(x_n \mid C_i) \quad (2.12)$$

$P(X|C_i)$ được tính với giả định x_k độc lập có điều kiện; $k = 1..n$:

- $P(x_k|C_i)$ được tính như sau:

+ Nếu *X* là các giá trị rời rạc

$$P(C_i) = \frac{|C_{i,D}|}{D}$$
$$P(x_k \mid C_i) = \frac{\#C_{i,D}\{x_k\}}{|C_{i,D}|} \quad (2.13)$$

+ Nếu *X* là các giá trị liên tục: $P(x_k|C_i)$ được ước lượng qua hàm mật độ:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (2.14)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{\frac{(x_k-\mu_{Ci})^2}{2\sigma_{Ci}^2}} \quad (2.15)$$
$$\mu = \frac{1}{n}\sum_{k=1}^{x} x_k \quad (2.16)$$

ở đây μ: Giá trị trung bình; σ: Độ lệch chuẩn.

$$\sigma = \frac{1}{n-1}\sum_{k=1}^{x}(x_{k-\mu})^2 \quad (2.17)$$

Tóm lại, để phân lớp mẫu *X*, tính: $P(X|C_i)$ $P(C_i)$ cho từng C_i, gán *X* vào lớp C_i sao cho $P(X|C_i)$ $P(C_i)$ là lớn nhất.

$$\max_{C_i \in C}(P(C_i \prod_{k=1}^{n} P(x_k \mid C_i))) \quad (2.18)$$

2.4. Thử nghiệm

Bài báo thử nghiệm bài toán trên phần mềm Weka với dữ liệu giống với dữ liệu sử dụng thuật toán ID3. Sau khi thử nghiệm bài toán trên phần mềm Weka sử dụng thuật toán Bayes chúng ta được kết quả như sau:

```
=== Summary ===

Correctly Classified Instances          81               72.973 %
Incorrectly Classified Instances        30               27.027 %
```

Hình 3. *Kết quả xác nhận phân lớp Bayes*

Kết quả dự đoán thuật toán NavieBayes với dữ liệu trong phần mềm Weka, với kết quả dự đoán đúng 81 giá trị chiếm 72,973%, kết quả dự đoán sai 30 giá trị chiếm 27,027%.

3. Các độ đo đánh giá

Trong học máy và trong các bài toán phân lớp theo thống kê), ma trận nhầm, còn gọi là ma trận lỗi (*Error Matrix*) hoặc ma trận so khớp (*Matching Matrix*) là bảng vuông, hai chiều cho phép thể hiện trực quan các chỉ tiêu đánh giá thuật toán cây quyết định cho lớp bài toán phân lớp, dự đoán (Predict). Trong ma trận, mỗi dòng mô tả các trường hợp xảy ra (*Instances*) theo thực tế (*Actual Class*); mỗi cột thể hiện các trường hợp xảy ra theo dự đoán (*Predicted Class*) hoặc ngược lại khi bảng đổi dòng thành cột (hay ma trận được chuyển vị). Trường hợp số lớp là *n, ma trận nhầm lẫn sẽ là nxn*. Trong đó:

- *Số mẫu dương Condition Positive (P);*
- *Số mẫu âm (Condition Negatives (N);*
- *Thực dương TP (True Positive);*
- *Thực âm TN (True Negativeeqv. with Correct Rejection)*
- *Dương sai FP (False Positiveeqv. with False Alarm, Type I Error:* báo động sai, sai số loại I);
- *Âm sai FN (False Negative eqv. with Miss, Type II Error*: mất, sai số loại II).

Bảng 2. *Ma trận nhầm lẫn và các chỉ tiêu đánh giá*

		predicted condition			
	total population	prediction positive	prediction negative	Prevalence = $\frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$	
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection = $\frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	False Negative Rate (FNR), Miss Rate = $\frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm = $\frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	True Negative Rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
Accuracy = $\frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$		Positive Predictive Value (PPV), Precision = $\frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) = $\frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False Discovery Rate (FDR) = $\frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) = $\frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

Các chỉ tiêu đánh giá

1. Tỷ lệ thực dương TPR (True Positive Rate) hay độ nhạy (Sensitivity) hay tỷ lệ trúng đích (Hit Rate) hay độ thu hồi (Recall)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (3.1)$$

2. Tỷ lệ thực âm TNR: (True Negative Rate) hay độ đặc hiệu SPC: (Specificity)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (3.2)$$

3. Giá trị dự đoán dương PPV: (Positive Predictive Value), hay giá (Precision)

$$PPV = \frac{TP}{TP + FP} \quad (3.3)$$

4. Giá trị đoán âm NPV: (Negative Predictive Value)

$$NPV = \frac{TN}{TN + FN} \quad (3.4)$$

5. Tỷ lệ bỏ lỡ hoặc tỷ lệ sai âm (FNR)

$$FNR = \frac{FN}{N} = \frac{FN}{FN + TP} = 1 - TPR \quad (3.5)$$

6. Rơi ra hoặc tỷ lệ dương giả (FPR: Fall-out or False Positive Rate)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR \quad (3.6)$$

7. Tỷ lệ khám phá sai (FDR: False Discovery Rate)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV \quad (3.7)$$

8. Tỷ lệ bỏ sót sai (FOR)

$$FOR = \frac{FN}{FN + TN} = 1 - NPV \quad (3.8)$$

9. Độ chính xác (ACC:accuracy)

$$ACC = \frac{TP + TN}{P + N} \quad (3.9)$$

10. Điểm số F1: Là trung bình hài của độ chính xác và độ nhạy (Score is the Harmonic Mean):

$$F_1 = 2 . \frac{PPV . TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN} \quad (3.10)$$

11. Hệ số tương quan Matthews (MCC: Matthews Correlation Coefficient)

$$MCC = \frac{TP . TN - FP . FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.11)$$

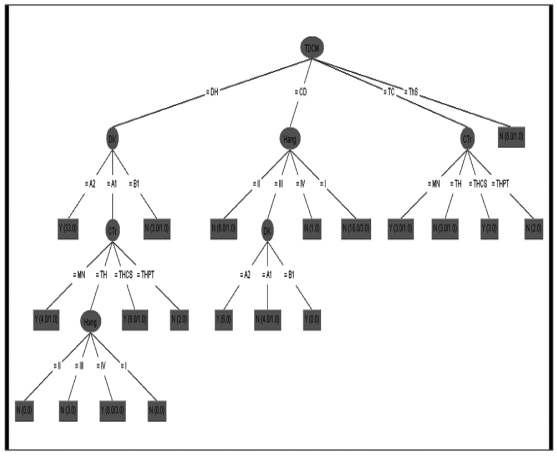
12. Chỉ tiêu BM (Bookmaker Informedness)

$$BM = TPR + TNR - 1 \quad (3.12)$$

13. Đánh dấu (MK: Markedness)

$$MK = PPV + NPV - 1 \quad (3.13)$$

Sơ đồ cây cho bài toán phân chia lớp sau khi sử dụng phần mềm Weka-6.6 cho bài toán phân chia lớp ngoại ngữ dựa trên thuật toán C4.5 (J48).

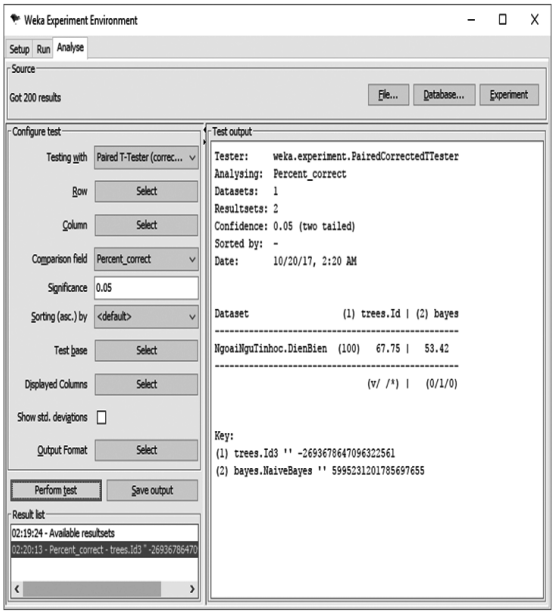


Hình 3. *Sơ đồ cây chia lớp ngoại ngữ*

4. So sánh kết quả và kết luận

4.1. So sánh thuật toán ID3 và Bayes

Để so sánh và phân tích kết quả thử nghiệm của hai thuật toán ID3 và Bayes ta sử dụng phần mềm Weka cho kết quả như Hình 4.



Hình 4. *Giao diện so sánh ID3 và Bayes*

Từ Hình 4 ta thấy, thuật toán ID3 cho kết quả dự đoán đúng là 77,48% trong khi thuật toán Bayes cho kết quả 72,97%. Điểm giống nhau giữa ID3 và Bayes:

+ Hai phương pháp đều là mô hình học có giám sát, dạng cây (cho trước đầu ra là nhãn); cần có tập dữ liệu mẫu huấn luyện.

Điểm khác nhau giữa ID3 và Bayes:

+ Thuật toán ID3 xây dựng mô hình cây với các nút lá được gán nhãn và rút ra các tập luật *if-then* tương ứng.

+ Thuật toán Bayes ước lượng xác suất của các lớp đã được gán nhãn thông qua dữ liệu huấn luyện và các đặc trưng đầu vào để gán nhãn cho các mẫu mới.

4.2. Kết luận

Đóng góp chủ yếu của bài báo là thu thập xử lý dữ liệu, thử nghiệm phân chia học viên đăng ký học Ngoại ngữ, phân loại cán bộ giáo viên trường nghề sử dụng thuật toán ID3, C4.5 và Bayes bằng tính toán trực tiếp và bằng phần mềm Weka cho một số kết quả khả quan; có thể ứng dụng cho các trường tương tự. Căn cứ kết quả đó, Trung tâm sẽ xử lý thông tin chính xác, nhanh bằng phần mềm về chia lớp học cho học viên đăng ký học Ngoại ngữ để cho có hiệu quả hơn.

Hướng nghiên cứu tiếp theo:

Sẽ thử nghiệm bài toán với khối lượng mẫu lớn hơn để đánh giá độ tin cậy của các thuật toán phân lớp học viên.

Lời cảm ơn

Bài báo được hỗ trợ từ trường Đại học Sư phạm Kỹ thuật Hưng Yên theo nội dung của nhóm nghiên cứu "Tính toán mềm" được Quyết Định số 1417/QĐ-ĐHSPKTHY ngày 06/07/2017.

Tài liệu tham khảo

[1]. Trần Cao Đệ, Phạm Nguyên Khang (2012), *Phân loại văn bản với máy học Vector hỗ trợ và cây quyết định*, Tạp chí Khoa học 2012:21a 52-63, Đại học Cần Thơ.

[2]. Nguyễn Quang Hoan (2007), *Nhập môn trí tuệ nhân tạo*, Học viện Công nghệ Bưu chính Viễn thông.

[3]. Nguyễn Dương Hùng (2000), *Hạn chế rủi ro tín dụng dựa trên thuật toán phân lớp*, Khoa Hệ thống Thông tin Quản lý – Học viện Ngân hàng.

[4]. Đỗ Thanh Nghị (2008), *Phương pháp K láng giềng - K Nearest Neighbors*, Khoa Công nghệ Thông tin – Đại học Cần Thơ.

[5]. Đỗ Thanh Nghị (2008), *Phương pháp học Bayes - Bayesian classification*, Khoa Công nghệ thông tin – Đại học Cần Thơ.

[6]. Võ Văn Tài (2012), *Phân loại bằng phương pháp Bayes từ số liệu rời rạc*, Tạp chí Khoa học 2012:23b 69-78, Đại học Cần Thơ.

[7]. Andrew Colin (1996), *Building Decision Trees with the ID3 Algorithm*, Dr. Dobbs Journal.

[8]. ShwetaKharya, SunitaSoni (2016), *Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection*, International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016, Bhilai Institute of Technology, Durg C.G. India.

[9]. Megha Gupta, Naveen Aggarwal (2010), *Classification Techniques Analysis*, UIET Punjab University Chandigarh INDIA -160014.

[10]. Miss. Deepa S. Deulkar& Prof .R. R. Deshmukh (2016), *Data Mining Classification*, Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-4, 2016 ISSN: 2454-1362, H.V.P.M. COET, Amaravati, India.

DECISION TREE ALGORITHMS AND CLASSIFIER EVALUATION BY CONFUSION MATRIX

Abstract:

The paper analyzed ID3, C4.5, Bayesalgorithms and we were coding data to classify. The ID3, C4.5, Bayesalgorithms are used to classify for English Leaners, for staff of Dienbien Vocational College. The paper proposed the criteria for classifier evaluation by confusion matrix to evaluate the classifier results.

Keyworks: *InformationGain, Machine Learning, ID3-Algorithm, BayesAlgorithm.*