ISSN 2354-0575

CÁC THUẬT TOÁN HỌC MÁY ÁP DỤNG CHO PHÂN LỚP, DỰ ĐOÁN

Nguyễn Quang Hoan[1], Vũ Ngọc Tân[2], Nguyễn Bá Giầu[3], Phạm Đình Hà[4]
1 Trường Đại học Sư phạm Kỹ thuật Hưng Yên
2 Trường Đại học Luật Hà Nội
3 Trường THPT Nguyễn Bỉnh Khiêm - Vĩnh Bảo - Hải Phòng
4 Trường THPT Kim Thành - Kim Thành - Hải Dương

Ngày tòa soạn nhận được bài báo: 12/01/2018
Ngày phản biện đánh giá và sửa chữa: 26/02/2018
Ngày bài báo được xét duyệt đăng: 28/02/2018

Tóm tắt:

Bài báo nghiên cứu các thuật toán C4.5, Bayes áp dụng cho các bài toán phân lớp và dự báo. Các chỉ tiêu theo ma trận nhầm lẫn được sử dụng để đánh giá, so sánh hiệu quả của các thuật toán. Một số bàn luận các trường hợp khác nhau về độ lớn dữ liệu, tỷ lệ dữ liệu học và dữ liệu thử được trình bày nhằm phân tích các tình huống trong ứng dụng. Các tính toán trực tiếp so sánh với tính toán trong phần mềm Weka được sử dụng để chạy thử, kết xuất, hiển thị các kết quả phân lớp, dự báo nhằm minh chứng khả năng áp dụng thực tế.

Từ khóa: *Luật học; Độ lợi thông tin, Entropy, thuật toán học, tỷ suất độ lợi.*

1. Giới thiệu

Cây quyết định là một phương pháp tạo ra cấu trúc, trong đó mỗi nút đại diện cho một "phép thử" đối với một thuộc tính; mỗi nhánh trong cây thể hiện một kết quả thu được trên cơ sở các luật [4, 6]. Đường dẫn từ gốc đến lá đại diện cho quy trình trong phân loại. Cây quyết định dựa trên phương pháp "*chia (nói theo hình tượng là cây) để trị*". "*Trị*" ở đây hàm ý rút ra các quy tắc, các luật học. Luật phổ biến là luật *if...then* (hay *luật nhân quả*) được áp dụng.

Có hai cách chia toàn bộ cơ sở dữ liệu học thành cây, đó là: chia theo các đặc trưng đầu vào và chia theo các đặc tính hoặc giá trị đầu ra. Chia theo các đặc trưng đầu vào có nhiều kỹ thuật chọn gốc khác nhau tùy theo luật học như: theo véc tơ xác suất xuất hiện các giá trị của đặc trưng, điển hình là thuật toán Quinlan; chia dùng tiêu chuẩn Entropy có bốn phiên bản điển hình: thuật toán độ lộn xộn, thuật toán ID3, thuật toán C4.5, thuật toán C5.0 [1, 2, 3, 6, 7, 8]. Chia theo đầu vào dùng mạng nơ ron nhân tạo có các thuật toán học như Perceptron, lan truyền ngược, Hebb… Chia theo đầu vào, sử dụng xác suất có điều kiện có thuật toán Naïve Bayes [1], với các giả thiết các đặc trưng đầu vào độc lập với nhau; hoặc dùng mạng Bayes [4], khi có đủ dữ liệu về xác suất có điều kiện. Chia theo đầu ra có thuật toán học quy nạp (Inductive Learning Algorithm) [3]. Thuật toán C4.5 phù hợp với các cơ sở dữ liệu vừa và nhỏ nên ứng dụng khá phổ biến và được chọn dùng trong bài báo này. Cùng với các thuật toán cây quyết định, ma trận nhầm lẫn với các chỉ tiêu đánh giá hiệu quả phân lớp dữ liệu [8] được áp dụng. Bài báo này còn nêu quy trình giải bài toán phân lớp dựa theo cây quyết định, cách lựa chọn thuật toán và cách xử lý dữ liệu cho các đối tượng cụ thể.

2. Thuật toán cây quyết định

2.1. Thuật toán C4.5

C4.5 được Breiman, Friedman, Olsen và Stone phát triển từ thuật toán ID3 trong lĩnh vực trí tuệ nhân tạo và trong thống kê. ID3 sử dụng độ lợi thông tin (*Information Gain*) làm tiêu chí chọn nút. Độ lợi thông tin của một thuộc tính được tính bằng độ đo hỗn loạn trước khi phân hoạch trừ cho độ đo hỗn loạn sau khi phân hoạch. Gọi S là tập dữ liệu huấn luyện; $C_{i,S}$: tập con (hay đặc tính) các mẫu thứ i của S; $i=\{1,...,k\}$, k: số tập con; $|C_{iS}|,|S|$: là số lượng các mẫu (hay lực lượng) của tập $C_{i,S}$ và S một cách tương ứng; p_i là xác suất xuất hiện số mẫu thuộc lớp C_i so với tổng số mẫu. Độ đo hỗn loạn thông tin trước khi chia tập con (phân hoạch) được tính theo:

$$Info\ (S) = -\sum_{i=1}^{k} p_i \log_2 p_i \qquad (2.1)$$

Độ đo hỗn loạn sau khi phân hoạch S thành k phần được tính:

$$Info_A(S) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Info(S_i) \qquad (2.2)$$

Độ lợi thông tin (*Information Gain*):

$$Gain(S) = Info(S) - Info_A(S) \qquad (2.3)$$

Khi dữ liệu có một thuộc tính chứa nhiều giá trị hơn các thuộc tính khác, độ lợi thông tin tăng trên các thuộc tính đó. Để giảm bớt sự chênh lệch này, Quinlan [6] sử dụng tỉ số độ lợi *Gain Ratio*. Tỉ số độ lợi được tính bằng độ lợi thông tin chia cho Entropy

phân phối dữ liệu trên các nhánh: *SplitInfo*(*S*).

$$SplitInfo(S) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (2.4)$$

Và tỉ số độ lợi:

$$GainRatio(S) = \frac{Gain(S)}{SplitInfo(S)} \quad (2.5)$$

Trong C4.5 dùng chỉ số *Information Gain* hoặc *Gain ratio* để xác định thuộc tính tốt nhất làm tiêu chí chọn gốc, trong đó *Gain Ratio* là lựa chọn mặc định.

2.2. Mô hình dự báo Bayes

Mô hình dự báo Bayes dựa trên định lý xác suất Bayes như sau [1]:

$$P[H|E] = \frac{P[E|H]P[H]}{P[E]} \quad (2.6)$$

trong đó, tập đặc trưng đầu vào $E = [E_1, E_2, \ldots, E_n]$ có n thuộc tính sử dụng làm bằng chứng (*Evidences*) và H là giả thuyết hay "nhãn" đầu ra cần dự báo.

Với giả thiết các đặc trưng đầu vào độc lập nhau, cách dự báo trong mô hình Bayes đơn giản được tính [2]:

$$P[H|E] = \frac{P[E_1|H].P[E_2|H]\cdots P[E_n|H]P[H]}{P[E]} \quad (2.7)$$

Các bước thực hiện thuật toán Bayes

Bước 1: Từ tập dữ liệu huấn luyện *S*, tính xác suất của các lớp;

Bước 2: Phân lớp $x=(x_1,\ldots,x_n)$, tính xác suất của các lớp đã xác định.

Bước 3: Tính (2.7) với các giá trị của *H*. Sử dụng quy luật "số lớn": với giá trị nào của *H* theo (2.7) lớn nhất, gán kết quả tương ứng cho giá trị đó của xác suất điều kiện trên tập dữ liệu *S*.

3. Xử lý dữ liệu và phân lớp

Đặt bài toán: Trong những năm gần đây, để nhập trường phổ thông trung học (THPT) nào đó, học sinh khi biết điểm thi tốt nghiệp Phổ thông Cơ sở có thể dự đoán có thể vào trường nào. Chúng tôi xây dựng các bước tạo cơ sở dữ liệu huấn luyện, sau đó phân lớp và dự báo. Tiến trình được mô tả trong bài báo dựa trên ví dụ ở trường trung học phổ thông (THPT) Nguyễn Bỉnh Khiêm, Hải Phòng làm minh họa và triển khai, áp dụng tượng tự được cho các trường khác như THPT Kim Thành, Hải Dương và cả đại học Luật. Quá trình đó được mô tả và thực hiện như sau.

3.1. Thu thập, phân tích dữ liệu

Dưới đây là một số căn cứ để xây dựng cơ sở dữ liệu học sinh theo bài toán.

- Căn cứ thông tư số 11/2014/TT-BGDĐT ngày 18/04/2014 của Bộ trưởng Bộ giáo dục và đào tạo (BGDĐT) ban hành quy chế thi trung học phổ thông (THPT); Thông tư số 18/2014/TT-BGDĐT ngày 26/05/2014 của Bộ trưởng BGDĐT, bổ sung kèm theo thông tư số 11/2014/TT-BGDĐT ngày 18/04/2014 của Bộ trưởng BGDĐT; Thông tư số 06/2012/TT-BGDĐT ngày 15/02/2012 của Bộ trưởng BGDĐT ban hành quy chế tổ chức và hoạt động của trường THPT; Căn cứ thông tư số 12/2014/TT-BGDĐT ngày 18/04/2014 của Bộ trưởng BGDĐT sửa đổi, bổ sung Điều 23, điều 24 quy chế tổ chức và hoạt động của trường THPT kèm theo thông tư số 06/2012/TT-BGDĐT ngày 15/02/2012 của Bộ trưởng BGDĐT.

- Căn cứ văn bản số 3090/UBND-VX ngày 18/12/2016 của UBND thành phố Hải Phòng về phương án tuyển sinh vào 10 THPT năm học 2017-2018: hình thức thi tuyển gồm 3 bài thi môn: Toán, Ngữ văn, Bài thi tổ hợp (gồm các môn: Vật lý, Hóa học, Lịch sử, Địa lý, GDCD, Sinh học, Ngoại ngữ (tiếng Anh)). Các bài thi đều hệ số điểm 20.

3.2. Xử lý, phân tích dữ liệu

Dựa vào các thông số quy ước (Bảng 3.1), điểm thi vào trường THPT Nguyễn Bỉnh Khiêm năm học 2017- 2018, trích rút từ 375 học sinh (dữ liệu gốc) và loại bỏ các trường hợp số liệu trùng nhau, quy đổi sang biến ngôn ngữ...tao tệp dữ liệu học (Bảng 3.2) với 15 mẫu (không trùng lặp). Chọn 5 đặc trưng đầu vào: "Điểm văn", Điểm toán"; "Điểm tổng hợp"; "Điểm ưu tiên"; "Tổng điểm" theo các quy chế đã nêu. Mỗi đặc tính đầu vào có thể nhận nhiều giá trị khác nhau với thang điểm 20. Về nguyên tắc, có thể chọn 20 giá trị khác nhau, nhưng với bối cảnh bài toán, chúng tôi chuyển dữ liệu gốc về bốn giá trị: "khá", "giỏi", "trung bình", "kém"; Đặc tính tổng điểm chọn hai giá trị "Đạt", "Không đat" đủ thể hiện. Kết quả đầu ra nhận 2 giá trị "Đỗ", "Không đỗ" ứng với giá trị nhị phân "Y', "N" (Bảng 3.1). Sau khi mã hóa, ta có 5 đặc trưng vào, nhan (đầu ra) có hai giá trị như Bảng 3.2.

Bảng 3.1. *Quy ước biểu diễn dữ liệu*

Giá trị	Ý nghĩa	Giá trị điểm
VGK	**văn** giỏi-khá	14>=VGK<= 20
VTB	**văn** trung bình	7 >= VTB < 14
VY	**văn** yếu	VY <7
TGK	**toán** giỏi-khá	14 >=TGK <= 20
TTB	**toán** trung bình	7 >= TTB < 14
TY	**toán** yếu	TY <7
THGK	**Tổ hợp** giỏi khá	14 >=THGK<= 20
THTB	**Tổ hợp** trung bình	7 >=THTB < 14

THY	**Tổ hợp** yếu	THY <7
UTG	**Ưu tiên** giỏi	>= 3 < 5
UTK	**Ưu tiên** khá	>= 1.5< 3
UTTB	**Ưu tiên** trung bình	<1.5 >0
T	**Tông điểm** Đạt	T >=30
F	**Tổng điểm** Không đạt	F < 30
Y		HS đỗ
N		HS không đỗ

3.3. Phân lớp với C4.5 và đánh giá

Dựa trên dữ liệu học đã được xác định (Bảng 3.2) sử dụng các công thức 2.1-2.5, chúng tôi thu được kết quả như Bảng 3.3 và đó là các chỉ tiêu chọn cây. Tiến hành tính toán theo các bước của thuật toán bằng tay và đồng thời thử nghiệm bằng phần mềm Weka, chúng tôi thu được kết quả tương đương. Các luật học được rút ra (Bảng 3.4). Chúng tôi thử nghiệm cho hai bộ dữ liệu: dữ liệu học (15 mẫu như Bảng 3.2) và dữ liệu gốc (375 học sinh).

Bảng 3.2. *Bảng dữ liệu học*

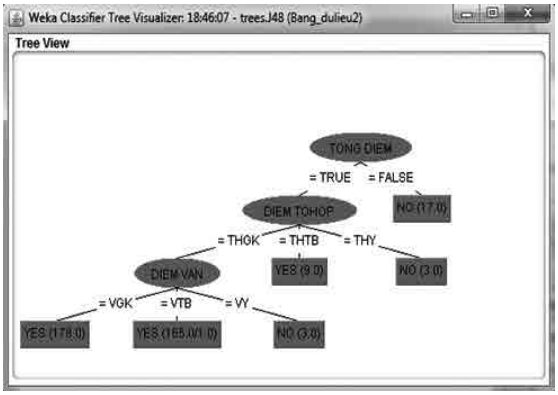
TT	Điểm Văn	Điểm Toán	Điểm tổng hợp	Điểm ưu tiên	Tổng điểm	Quyết định
1	VGK	TGK	THGK	UTG	T	Y
2	VGK	TGK	THGK	UTK	T	Y
3	VGK	TGK	THGK	UTTB	T	Y
4	VTB	TGK	THGK	UTG	T	Y
5	VTB	TGK	THGK	UTK	T	Y
6	VTB	TGK	THGK	UTTB	T	Y
7	VY	TGK	THGK	UTK	T	N
8	VY	TY	THTB	UTK	F	N
9	VGK	TTB	THGK	UTG	T	Y
10	VTB	TTB	THGK	UTK	T	Y
11	VTB	TY	THGK	UTK	T	N
12	VTB	TY	THTB	UTK	F	N
13	VGK	TY	THY	UTTB	F	N
14	VY	TTB	THGK	UTK	T	N
15	VY	TTB	THTB	UTK	F	N

Bảng 3.3. *Bảng thuộc tính, chỉ tiêu*

STT	THUỘC TÍNH	GAIN	SPITINFO	GAINRATIO
1	D.VAN	0.39	1.56	0.25
2	D.TOAN	0.45	1.53	0.29
3	DIEMTH	0.37	1.09	0.34
4	DIEMUT	0.22	1.34	0.16
5	**T.DIEM**	**0.62**	**0.84**	**0.74**

Bảng 3.4. *Tập luật cho S*

1	if (t.diem = F) then (ketqua = N)
2	if (*t.diem = T*) and (*diemvan = VY*) then (*ketqua = N*)
3	if (*t.diem = T*) and (*diemvan = GK*) then (*ketqua = Y*)
4	if (*t.diem = T*) and (*diemvan = VTB*) then (*ketqua = Y*)



Hình 3.1. *Cây phân lớp cho 375 học sinh*

Tiêu chuẩn đánh giá phân lớp quan trọng nhất (trong 13 tiêu chuẩn) là độ chính xác [1]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3.1)$$

trong đó, *TP*: True positive (thực dương), *TN*: True positive (thực âm); *FP*: False positive (Sai dương); *FN*: False negative (sai âm). Tỷ lệ học chọn 80%, tỷ lệ thử 20% của 375 là 75 người phân lớp đúng: 72 người, chiếm 96%, tỷ lệ phân lớp sai 3 người, chiếm 4%. Cây quyết định được cho trên Hình 3.1.

4. Dự báo tuyển sinh thuật toán Bayes

Bài toán: Giả sử học sinh tên "Hùng" có kết quả: "Điểm Văn"=VTB; "Điểm Toán"=TKG: "Điểm Tổng hợp" =THTB; "Điểm Ưu tiên"=UTG; "Tổng điểm"=T. Dự báo "Hùng" đỗ, trượt?

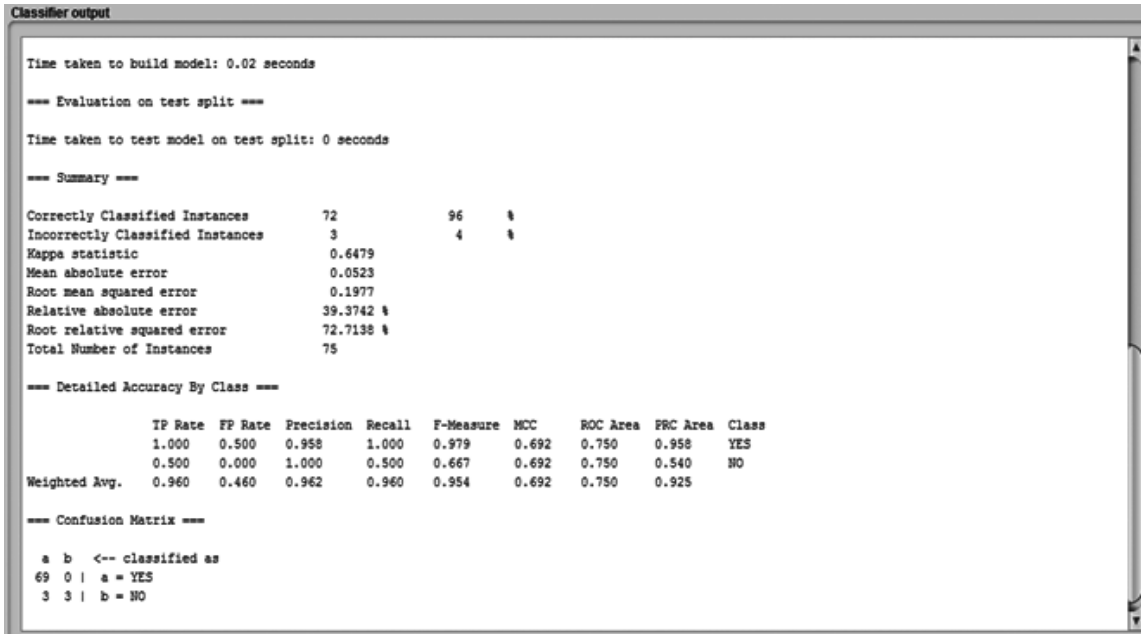
Theo (2.7) và các bước ở mục 2.2, dự báo được tính với hai "Quyết Định" (QĐ): *P(QĐ=Y|x), P(QĐ=N|x)* với *x={điểm toán, điểm văn, điểm tổng hợp, điểm ưu tiên, tổng điểm}*

• *P[QĐ=Y|x] = {P[Điểm văn=VTB|QĐ=Y].*
P[Điểm toán=TKG|QĐ=Y].
P[Điểm tổ hợp=THTB|QĐ=Y].
P[Điểm ưu tiên=UTG|QĐ=Y].
P[Tổng điểm=T|QĐ=Y]}/ P(x)=
=(4/8).(6/8).(0/8).(0/8).(8/8).(8/15)=0

(Do hai công thức trên có mẫu số đều là *P(x)*; để so sánh, chỉ cần tính tử số)

• *P[QĐ=N|x]={P[Điểm văn=VTB|QĐ=N].*
P[Điểm toán=TKG|QĐ=N].
P[Điểm tổ hợp=THTB|QĐ=N].
P[Điểm ưu tiên=UTG|QĐ=N].
P[Tổng điểm=T|QĐ=N]}/P(x)=
=(4/7).(1/7).(3/7).(3/7).(3/7).(7/15)= 0,003>*P[QĐ=Y|x]=0.->QĐ=N (trượt).*

Đánh giá thuật toán: Về định tính, thuật toán giả thiết các đặc trưng đầu vào độc lập nhau, điều đó dẫn tới độ chính xác kém khách quan. Về định lượng, hoàn toàn có thể đưa ra bảng chỉ tiêu giống như Hình 3.2 theo phần mềm Weka, hoặc tính toán với độ chính xác tương tự. Để có kết quả khả quan hơn có thể sử dụng mạng Bayes [4] nếu đủ số liệu về xác suất giữa các đặc trưng.



Hình 3.2. *Bảng kết quả phân lớp cho tệp dữ liệu gốc (375 học sinh)*

Bảng 4.1. *Bảng tính xác suất mẫu*

Điểm Văn			Điểm Toán			Điểm Tổ Hợp			Điểm Ưu Tiên			Điểm Tổng			Kết quả	
	Y	N		Y	N		Y	N		Y	N		Y	N	Y	N
VGK	4	1	TGK	6	1	THGK	8	3	UTG	0	3	T	8	3	8	7
VTB	4	2	TTB	2	2	THTB	0	3	UTK	5	4	F	0	4		
VY	0	4	TY	0	4	THY	0	1	UTTB	2	1					
P(VGK)	4/8	1/7	P(TGK)	6/8	1/7	P(THGK)	8/8	3/7	P(UTG)	0/8	3/7	P(T)	8/8	3/7	8/15	7/15
P(VTB)	4/8	2/7	P(TTB)	2/8	2/7	P(THTB)	0/8	3/7	P(UTK)	5/8	4/7	P(F)	0/8	4/7		
P(VY)	0/8	4/7	P(TY)	0/8	4/7	P(THY)	0/8	1/7	P(UTTB)	2/8	1/7					

5. Kết luận và hướng phát triển tiếp

Đóng góp của bài báo là xây dựng, xử lý dữ liệu, quy trình thực hiện phân lớp, dự báo số học sinh tuyển sinh của trường THPT Nguyên Bỉnh Khiêm, Hải Phòng; THPT Kim Thành, Hải Dương; cải biên áp dụng cho đại học Luật, Hà Nội (chỉ khác về các đặc trưng, giá trị đầu vào; thuật toán cơ bản giống nhau). Bài toán có thể cải biên và áo dụng cho các trường phổ thông và trường đại học khác.

Hướng phát triển tiếp theo là xây dựng phần mềm phân lớp và tra cứu điểm, đáp ứng công tác quy hoạch và tin học hóa cho các trường phổ thông và đại học. Về mặt khoa học và công nghệ, phương pháp dự báo dùng mạng Bayes có thể cho kết quả chính xác hơn dùng thuật toán Bayes. Ngoài ra, chúng tôi có ý định sử dụng hệ lai mạng nơ ron, logic mờ và thuật toán di truyền hy vọng cho kết quả dự báo chính xác hơn nữa và sẽ có báo cáo sau.

Nghiên cứu này được tài trợ bởi Trung tâm Nghiên cứu Ứng dụng Khoa học và Công nghệ, Trường Đại học Sư phạm Kỹ thuật Hưng Yên, đề tài mã số: UTEHY.T028.P1718.02.

Tài liệu tham khảo

[1]. Đỗ Thanh Nghị. *Khai mỏ dữ liệu,* NXB Đại học Cần Thơ, 2011.
[2]. Nguyễn Quang Hoan, Nguyễn Thị Thanh Lan, Hoàng Phú Quang, Phân loại chất lượng học sinh trường cao đẳng nghề xây dựng Quảng Ninh sử dụng phương pháp học máy. *Tạp chí Khoa học Công nghệ - Trường Đại học Sư phạm Kỹ thuật Hưng Yên*, ISSN 2354-0575, 2017, **số 14(3-2017)**, tr. 75-80.
[3]. Hoàng Kiếm, Đỗ Phúc, Đỗ Văn Nhơn. *Hệ cơ sở tri thức*, NXB Đại học Quốc gia Tp. Hồ Chí Minh, 2000.
[4]. Từ Minh Phương. *Trí tuệ nhân tạo*, NXB Thông tin và Truyền thông, 2016.
[5]. Trần Hoài Linh. *Mạng nơ-ron và ứng dụng trong xử lý tín hiệu*, NXB Bách khoa, Hà Nội, 2014.
[6]. Anurag Srivastava, Eui-Hong Han, Vipin Kumar, Vieet Singh. *Parallel Formulations of Decision-Tree Classification Algorithm*, Kluwer Academic Publisher, 1998.
[7]. Richard Kufrin, Generating C4.5 Production Rules in Parallel. In *Proceeding of Fourteenth National Conference on Artificial Intelligence, Providence RI*, 1997. doc.edu.vn/tai-lieu/nghien-cuu-cac-thuat-toan-phan-lop-du-lieu-dua-tren-cay-quyet-dinh-22489.
[8]. Ron Kohavi, J. Ross Quinlan, Data Mining Tasks and Methods: Classification: Decision-Tree Discovery. *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, Inc. New York, NY, USA ©2002, pp. 267-276.
[9]. The Morgan Kaufmann Series in Data Management Systems, Jim Gray. *Data Mining- Concepts and Techniques*, Chapter 7- Classification and Prediction. Series Editor Morgan Kaufmann Publishers, August, 2000.
[10]. Wu X. and Kumar V., *Top 10 Algorithms in Data Mining*, Chapman & Hall/CRC, 2009.

MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION, PREDICTION

Abstract:
This paper analyzed C4.5, Bayes algorithms for classification and the prediction problems. The classification criteria based on the confusion matrix are used to evaluate the classifier and predicted results. Weka software program was used to test the proposed classifier and predicted data.
Keywords: *Learning Rule, Information Gain, Entropy, Learning Algorithms, Gain Ratio.*