ISSN 2354-0575

PHÂN LOẠI CHẤT LƯỢNG HỌC SINH TRƯỜNG CAO ĐẲNG NGHỀ XÂY DỰNG QUẢNG NINH SỬ DỤNG PHƯƠNG PHÁP HỌC MÁY

Nguyễn Quang Hoan[1], Nguyễn Thị Thanh Lan[2], Hoàng Phú Quang[3], Đào Minh Tuấn[1]

1 Trường Đại học Sư phạm Kỹ thuật Hưng Yên

2 Trường Cao đẳng Nghề Xây dựng Quảng Ninh

3 Trường Cao đẳng Nghề Lạng Sơn

Ngày tòa soạn nhận được bài báo: 19/03/2017

Ngày phản biện đánh giá và sửa chữa: 22/05/2017

Ngày bài báo được duyệt đăng: 25/05/2017

Tóm tắt:

Phân lớp dữ liệu là một trong những hướng nghiên cứu của khai phá dữ liệu. Bài báo này phân tích, đánh giá và so sánh hai thuật toán tiêu biểu ID3 và Bayes trong phân lớp dữ liệu. Bài báo cũng tiến hành thử nghiệm hai thuật toán này cho bài toán phân loại chất lượng học sinh và cài đặt trên phần mềm Weka. Bài báo hướng tới cải thiện công tác quy hoạch và kế hoạch hóa bằng một số phần mềm nhằm nâng cao chất lượng dạy và học trong các trường nghề hiện nay.

Từ khóa: *Học máy, phân loại, cây quyết định, thuật toán ID3, thuật toán Bayes.*

1. Giới thiệu

Phân lớp dữ liệu đã, đang và sẽ phát triển mạnh mẽ trước những khao khát tri thức của con người. Trong những năm qua, phân lớp dữ liệu đã thu hút sự quan tâm các nhà nghiên cứu trong nhiều lĩnh vực khác nhau như học máy (Machine Learning), hệ chuyên gia (Expert System), thống kê (Statistics)... Công nghệ này cũng ứng dụng trong nhiều lĩnh vực thực tế như: thương mại, ngân hàng, Maketing, quản lý các đối tượng… nhằm hạn chế những rủi ro gặp phải [3]

Trong các mô hình phân lớp, thuật toán phân lớp là công cụ chủ đạo. Do vậy, chúng ta cần xây dựng những thuật toán có độ chính xác cao, thực thi nhanh, kèm với tính mở để có thể thao tác với những kho dữ liệu lớn (Big Data).

Bài toán đặt ra ở đây là phân loại học sinh của một trường nghề sử dụng phương pháp học máy. Có rất nhiều thuật toán phân lớp đã được công bố như: Cây quyết định (Thuật toán Quinlan, ID3, Độ lộn xộn, C4.5, C5.0…), K-NN (K-Nearest Neighbor), Bayes; học theo mạng nơron, hệ mờ… Mỗi thuật toán có ưu điểm, hạn chế và độ phức tạp khác nhau và được áp dụng cho nhiều lớp đối tượng [10]. Phương pháp 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 K-NN không tiến hành quá trình học, khi phân loại tốn nhiều thời gian do quá trình tìm kiếm k dữ liệu lân cận, kết quả phụ thuộc vào việc chọn khoảng cách và được ứng dụng nhiều trong lĩnh vực tìm kiếm thông tin, nhận dạng [4], [9]; thuật toán Bayes đơn giản cho kết quả tốt trong thực tế, mặc dù chịu giả thiết về tính độc lập xác suất giữa các thuộc tính và thường được ứng dụng trong các bài toán dự đoán, phân loại, phát hiện thư rác (Spam) [5], [8]. Các luật học dựa trên các mạng nơron nhân tạo chủ yếu ứng dụng trong các lĩnh vực nhận dạng, xử lý tiếng nói, điều khiển hay trong các lĩnh vực công nghệ thông tin, viễn thông [2] với độ phức tạp thuật toán cao.

Trong bài báo này, chúng tôi sử dụng thuật toán ID3 và Bayes để phân lớp chất lượng của học sinh Trường Cao đẳng Nghề Xây dựng Quảng Ninh. Đây là phương pháp phân loại mà từ trước tới nay trường chưa đưa vào sử dụng.

2. Các thuật toán chọn dùng cho phân loại chất lượng học sinh

Từ các phân tích trên, chúng tôi nhận thấy với quy mô dữ liệu không lớn, độ chính xác không đòi hỏi cao đối với một trường nghề có thể dùng thuật toán ID3 và Bayes để cho phân loại chất lượng học sinh.

2.1. Thuật toán ID3

Đầu vào: Cho tập dữ liệu huấn luyện gồm các thuộc tính A mô tả các tình huống, hay đối tượng nào đó, và một giá trị nhãn làm dấu hiệu để phân loại tình huống hoặc đối tượng đó.

Đầu ra: Cây quyết định đưa ra các luật có khả năng phân loại đúng các ví dụ mẫu trong tập dữ liệu đã được 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.

Thuật toán: Bắt đầu với nút gốc[1], [7]:

Bước 1: Chọn *A* ← thuộc tính quyết định "tốt nhất" cho *nút* kế tiếp.

Bước 2: Gán *A* là thuộc tính quyết định cho

nút.

Bước 3: Với mỗi giá trị của *A*, tạo nhánh con mới của nút.

Bước 4: Phân loại các mẫu huấn luyện cho các nhánh.

Bước 5: các mẫu huấn luyện trong một nhánh được phân loại hoàn toàn (đồng nhất một loại) thì DỪNG, ta được một nút lá; ngược lại, lặp với các nút nhánh mới.

Tiêu chí để chọn các gốc của thuật toán ID3 là độ lợi thông tin (Information Gain), được tính theo Entropy.

Độ lợi thông tin (Information Gain)

Tập dữ liệu *S* gồm có *n* thuộc tính A_i (i = 1, 2,…, n) độ lợi thông tin của thuộc tính *A* trong tập *S* ký hiệu là *Gain*(*S, A*) và được tính theo:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

+ Entropy của một tập *S*, có 2 lớp (nhị phân) dương (+) và âm (-) được tính:

$$Entropy(S) = - p_+ log_2 p_+ - p_- log_2 p_- \quad (2)$$

+ Entropy của tập *S* có *c* phân lớp (*c* nguyên, dương) có dạng tổng quát

$$Entropy(S) = \sum_{i=1}^{c} -P_i \log_2 P_i \quad (3)$$

trong đó, p_i: xác suất của các sự kiện đạt giá trị *i*, thuộc tập *S*.

2.1.1. Thử nghiệm bài toán bằng ID3

a. Phân tích và đặt bài toán

Trong bài báo, chúng tôi giới hạn 4 tham số (đặc trưng) chính ảnh hưởng đến chất lượng học sinh là: Xếp loại học lực, xếp loại đạo đức, Kỹ năng nghề và Tiếng Anh chuyên ngành. Mỗi yếu tố nhận các giá trị ngôn ngữ cụ thể như sau:

+ Biến 1: *Xếp Loại Học Lực (XLHL)*, có các giá trị: XLHL là "*Gioi*" khi điểm trung bình từ 8.0-9.0; XLHL là "*Kha*" khi điểm trung bình từ 7.0 – 8.0. XLHL là "*TBK*" khi điểm trung bình từ 6.0-7.0.XLHL là "*TB*" khi điểm trung bình từ 5.0-6.0.

+ Biến 2: *Xếp Loại Đạo Đức (XLDD)*: có các các giá trị: XLDD là "*Tot*" khi điểm rèn luyện từ 80-90 điểm; XLDD là "*DD_Kha*" khi điểm rèn luyện từ 70–80 điểm. XLDD là "*DD_TBK*" khi điểm rèn luyện từ 60–70 điểm.XLDD là "*DD_TB*" khi điểm rèn luyện từ 50–60 điểm.

+ Biến 3: *Kỹ năng nghề (KNN)*: Có hai giá trị: "*KNN_Tot*" và "*K_Tot*".

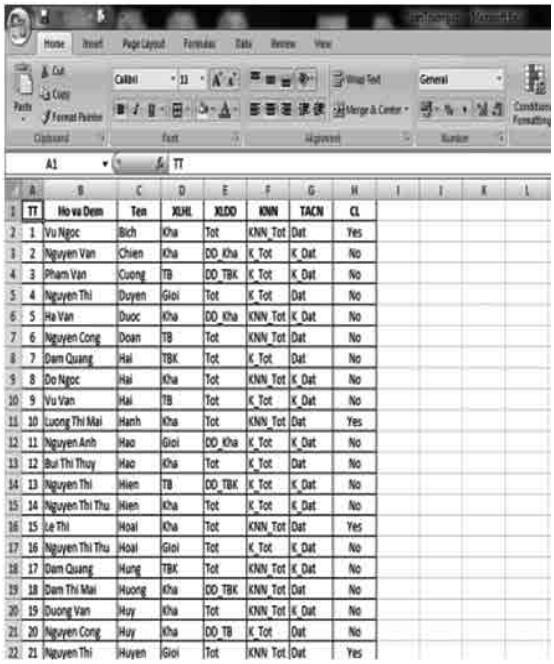
+ Biến 4: *Tiếng Anh chuyên ngành (TACN)*: Có hai giá trị: "*Dat*" và "*K_Dat*".

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* (Hình 1) chuyên nghiệp cho khai phá dữ liệu.

Bảng dữ liệu (Bảng1) với bốn thuộc tính: XLHL, XLDD, KNN, TACN và 650 bản ghi ứng với 650 học sinh trong toàn trường được lưu trong tệp *ToanTruong.CSV*

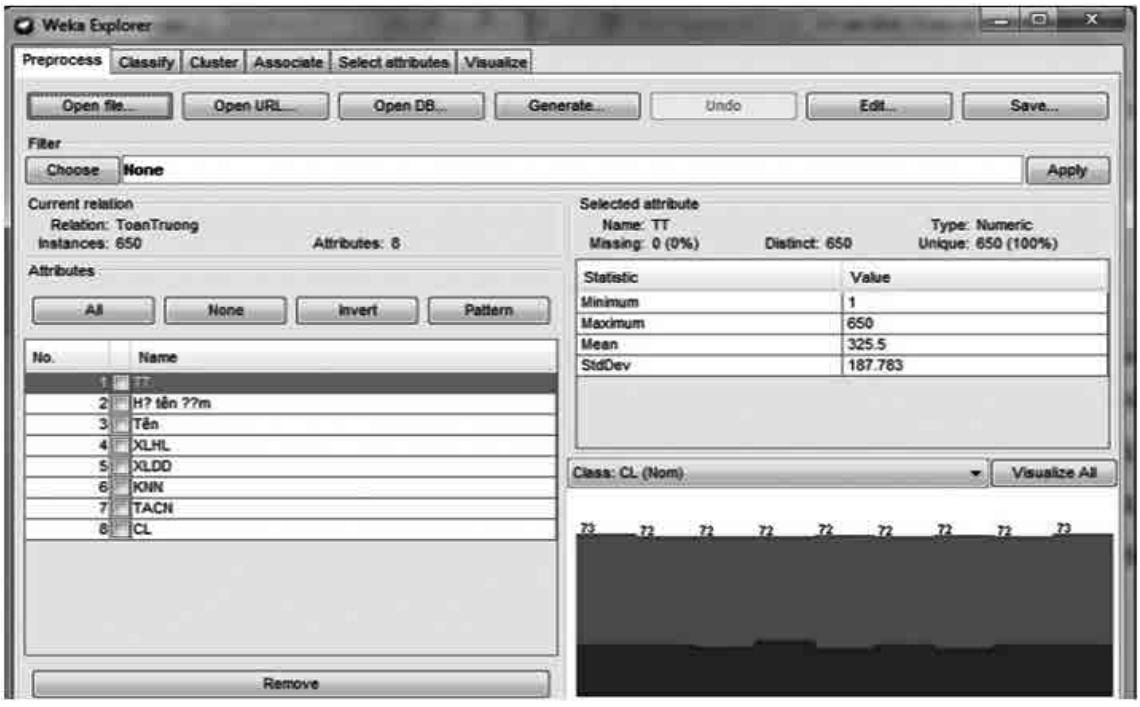
Bảng 1. *Tệp dữ liệu ToanTruong.CSV*



TT	Ho va Dem	Ten	XLHL	XLDD	KNN	TACN	CL
1	Vu Ngoc	Bich	Kha	Tot	KNN_Tot	Dat	Yes
2	Nguyen Van	Chien	Kha	DD_Kha	K_Tot	K_Dat	No
3	Pham Van	Cuong	TB	DD_TBK	K_Tot	K_Dat	No
4	Nguyen Thi	Duyen	Gioi	Tot	K_Tot	Dat	No
5	Ha Van	Duoc	Kha	DD_Kha	KNN_Tot	K_Dat	No
6	Nguyen Cong	Doan	TB	Tot	KNN_Tot	Dat	No
7	Dam Quang	Hai	TBK	Tot	K_Tot	Dat	No
8	Do Ngoc	Hai	Kha	Tot	KNN_Tot	K_Dat	No
9	Vu Van	Hai	TB	Tot	K_Tot	K_Dat	No
10	Luong Thi Mai	Hanh	Kha	Tot	KNN_Tot	Dat	Yes
11	Nguyen Anh	Hao	Gioi	DD_Kha	K_Tot	K_Dat	No
12	Bui Thi Thuy	Hao	Kha	Tot	K_Tot	Dat	No
13	Nguyen Thi	Hien	TB	DD_TBK	K_Tot	K_Dat	No
14	Nguyen Thi Thu	Hien	Kha	Tot	K_Tot	K_Dat	No
15	Le Thi	Hoai	Kha	Tot	KNN_Tot	Dat	Yes
16	Nguyen Thi Thu	Hoai	Gioi	Tot	K_Tot	K_Dat	No
17	Dam Quang	Hung	TBK	Tot	KNN_Tot	K_Dat	No
18	Dam Thi Mai	Huong	Kha	DD_TBK	KNN_Tot	Dat	No
19	Duong Van	Huy	Kha	Tot	KNN_Tot	K_Dat	No
20	Nguyen Cong	Huy	Kha	DD_TB	K_Tot	Dat	No
21	Nguyen Thi	Huyen	Gioi	Tot	KNN_Tot	Dat	Yes

Sau đó, ta tiến hành tiền xử lý dữ liệu với phần mềm Weka để chọn các thuộc tính cần thiết và loại bỏ các thuộc tính không cần thiết để phân loại (Hình 2, Hình 3):

Hình 1. *Giao diện chính phần mềm Weka*



Hình 2. *Các tham số trước khi lọc dữ liệu*

Hình 3. *Các tham số sau khi lọc dữ liệu*

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

```
Classifier output

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

Id3


XLDD = Tot
|  KNN = KNN_Tot
|  |  XLHL = Kha
|  |  |  TACN = Dat: Yes
|  |  |  TACN = K_Dat: No
|  |  XLHL = TB: No
|  |  XLHL = Gioi
|  |  |  TACN = Dat: Yes
|  |  |  TACN = K_Dat: No
|  |  XLHL = TBK: No
|  |  XLHL = XS
|  |  |  TACN = Dat: Yes
|  |  |  TACN = K_Dat: No
|  |  XLHL = Kem: No
|  KNN = K_Tot: No
XLDD = DD_Kha: No
XLDD = DD_TBK: No
XLDD = DD_TB: No
XLDD = DD_XS
|  XLHL = Kha
|  |  KNN = KNN_Tot
|  |  |  TACN = Dat: Yes
|  |  |  TACN = K_Dat: No
|  |  KNN = K_Tot: No
|  XLHL = TB: No
|  XLHL = Gioi
|  |  KNN = KNN_Tot
|  |  |  TACN = Dat: Yes
|  |  |  TACN = K_Dat: No
|  |  KNN = K_Tot: No
|  XLHL = TBK: No
|  XLHL = XS
|  |  KNN = KNN_Tot
|  |  |  TACN = Dat: Yes
|  |  |  TACN = K_Dat: No
|  |  KNN = K_Tot: No
|  XLHL = Kem: No
XLDD = DD_Kem: No
XLDD = Yeu: No
```

Hình 4. *Đầu ra: tập luật sử dụng ID3*

Kết quả: Khi sử dụng thuật toán ID3, ta rút ra được 27 tập luật (Hình 4) từ tập dữ liệu.

Bảng 2. *Kết quả phân lớp dùng ID3*

```
=== Summary ===

Correctly Classified Instances         638               98.1538 %
Incorrectly Classified Instances        12                1.8462 %
```

Hình 5. *Kết quả xác nhận phân lớp ID3*

2.1.2. Kết quả thử nghiệm dùng ID3

Bằng phương pháp tính tỷ số phần trăm ta có kết quả như Bảng 2:

- Trường hợp 1: Phân loại chính xác: 638 trường hợp, chiếm 98.1538%;

- Trường hợp 2: Phân loại không chính xác 12 trường hợp chiếm 1.8462%.

Bảng 3. *Ma trận nhầm lẫn dùng ID3*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 200   4 |   a = Yes
   8 438 |   b = No
```

Để đánh giá khả năng phân lớp, ta có thể sử dụng ma trận nhầm lẫn (Bảng 3) với kết quả như sau: có 4 trường hợp lớp *Yes* bị phân lớp nhầm sang lớp *No*; có 8 trường hợp lớp *No* bị phân lớp nhầm sang lớp *Yes*.

2.2. Thuật toán Naïve Bayes [5], [6]

Giả sử D là tập huấn luyện nhiều mẫu với vec tơ $X=(x_1, x_2,\ldots,x_n)$ và $C_{i,D}$ là tập các mẫu của D thuộc lớp C_i ($i = \{1,\ldots,m\}$). Các thuộc tính (x_1, $x_2,\ldots, x_n$) được giả thiết là độc lập nhau khi đó, xác suất có điều kiện Bayes được tính theo [5], [6]:

$$P = (C_i \mid X) = \prod_{i=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i) \quad (4)$$

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

- $P(x_k|C_i)$ được tính với hai trường hợp sau:

+ Nếu X là các giá trị rời rạc

$$P(C_i) = \frac{|C_{i,D}|}{D} \quad (5)$$

$$P(x_k \mid C_i) = \frac{\#C_{i,D}\{x_k\}}{|C_{i,D}|} \quad (6)$$

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

$$P(x_k \mid C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \quad (7)$$

$$\mu = \frac{1}{n}\sum_{k=1}^{x} x_k \quad (8)$$

trong đó, μ: Giá trị trung bình;

σ: Độ lệch chuẩn:

$$\sigma = \frac{1}{n-1}\sum_{k=1}^{x}(x_{k-\mu})^2 \quad (9)$$

Tóm lại, để phân lớp mẫu chưa biết X, ta 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.

$$\max_{C_i \in C}\left(P\left(C_i \prod_{k=1}^{n} P(x_k \mid C_i)\right)\right) \quad (10)$$

2.2.1. Thử nghiệm bài toán dùng Bayes

Bài báo cũng sử dụng thuật toán Bayes giải bài toán trên phần mềm *Weka* với cùng dữ liệu sử dụng cho thuật toán ID3; chúng ta có kết quả phân lớp đánh giá như sau:

Bảng 4. *Kết quả phân lớp dùng Bayes*

```
=== Summary ===

Correctly Classified Instances         648               99.6923 %
Incorrectly Classified Instances         2                0.3077 %
```

2.2.2. Kết quả thử nghiệm dùng Bayes

Từ Bảng 4 sử dụng phương pháp đánh giá theo phần trăm ta có:

- Phân loại chính xác: 648 trường hợp, chiếm 99.6923%;

- Phân loại không chính xác: 2 trường hợp chiếm 0.3077%.

Bảng 5. *Ma trận nhầm lẫn dùng Bayes*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 202   2 |   a = Yes
   0 446 |   b = No
```

Bảng 5, sử dụng phương pháp truyền thống: tính ma trận nhầm lẫn. Ta nhận thấy: có 2 trường hợp lớp *Yes* bị phân lớp nhầm sang lớp *No*; không có trường hợp nào lớp *No* bị phân lớp nhầm sang lớp *Yes*.

3. So sánh độ đo phân lớp của ID3, Bayes

Từ bảng ma trận nhầm lẫn ở Bảng 3 và Bảng 5 ta tính được các độ đo hiệu quả của việc phân lớp: **Precision, Recall, Accuracy** theo các công thức trong [5] cho hai thuật toán ID3 và Bayes như Bảng 6:

Bảng 6. *Các độ đo của thuật toán ID3, Bayes*

	Precision	**Recall**	**Accuracy**
ID3	0.9615	0.9803	0.9815
Bayes	1	0.9901	0.9969

Từ Bảng 6 ta thấy, với bài toán phân loại chất lượng học sinh Trường Cao đẳng Nghề Xây dựng Quảng Ninh sử dụng thuật toán Bayes sẽ có độ chính xác cao hơn khi sử dụng thuật toán ID3. Ngoài ra, ta có thể rút ra các đặc điểm của hai phương pháp:

Điểm giống nhau giữa ID3 và Bayes:

+ Cả hai phương pháp đều là mô hình học có giám sát, nghĩa là đều phải có một tập dữ liệu mẫu huấn luyện để chương trình có thể "học" qua ví dụ và rút ra các đặc trưng dùng cho việc gán nhãn.

+ Đều biết trước đầu ra: số nhãn.

Điểm khác nhau giữa ID3 và Bayes:

+ Thuật toán ID3 xây dựng cây quyết định 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 mẫu. Nếu xác suất của mẫu đó gần với giá trị đúng của lớp, ta gán mẫu cho lớp đó.

4. Kết luận

Đóng góp chủ yếu của bài báo là thử nghiệm phân loại chất lượng học sinh trường Cao đẳng Nghề Xây dựng sử dụng thuật toán ID3 và Bayes với một vài kết quả khả quan và có thể ứng dụng được cho các trường tương tự. Căn cứ kết quả đó, nhà trường sẽ có thông tin chính xác, nhanh bằng phần mềm về chất lượng học sinh để đưa ra các biện pháp dạy và học có hiệu quả hơn.

Hướng nghiên cứu tiếp theo:

Chúng tôi sẽ thử nghiệm bài toán với khối lượng mẫu lớn hơn để đánh giá độ tin cậy của các thuật toán trong phân loại học sinh của trường nghề.

Nghiên cứu, ứng dụng các thuật toán tiên tiến khác như C4.5 hay C5.0 thay cho ID3 để xử lý các trường hợp thiếu hoặc mất dữ liệu của các đặc trưng, nâng cao hiệu suất và tăng cường độ tối ưu cho ứng dụng.

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]. Deepa S. Deulkar, 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.

CLASSIFICATION OF THE STUDENT'S QUALITY IN THE QUANG-NINH BUILDING VOCATIONAL COLLEGE USING MACHINE LEARNING

Abstract:
Data classification is one of the major research areas of Data Mining. This paper is going to analyze, evaluate and compare two typical algorithms in data classifiers: ID3 and Bayes algorithms. Next, the article applies these algorithms for classifying the student's quality in the Quang-Ninh building vocational college using Weka software. This is a application in order to help for evaluating the quality of teaching and learning in the vocational college today.
Keyworks: *Machine Learning, Classification, Decision Tree, ID3Algorithm, BayesAlgorithm.*