

ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỂ PHÂN TÍCH DỮ LIỆU TUYỂN SINH DỰA VÀO XÉT ĐIỂM HỌC BẠ TẠI TRƯỜNG ĐẠI HỌC PHÚ YÊN TRONG NĂM HỌC 2018 – 2019

Hồ Thị Duyên*, Lê Thị Kim Anh
Trường Đại học Phú Yên

Tóm tắt

Trong bài báo này, chúng tôi giới thiệu ứng dụng khai phá dữ liệu để phân tích dữ liệu tuyển sinh dựa vào xét học bạ tại Trường Đại học Phú Yên (ĐHPY). Chúng tôi tiến hành sưu tập dữ liệu tuyển sinh năm học 2018 – 2019 tại Trường ĐHPY, sau đó thực hiện bước xây dựng cơ sở dữ liệu, tiền xử lý dữ liệu, đưa dữ liệu ra cấu trúc bảng. Chúng tôi đề xuất sử dụng giải thuật cây quyết định học từ dữ liệu để rút trích các tập luật quan trọng liên quan đến việc tuyển sinh. Kết quả thu được sau khi rút trích có thể cung cấp thông tin hữu ích cho việc tuyển sinh cao đẳng đại học và tổ chức kế hoạch tuyển sinh cho năm học sau.

Từ khóa: Khai phá dữ liệu, giải thuật cây quyết định, tuyển sinh

Abstract

Application of data using to analyze enrollment data based on academic records at Phu Yen University in the academic year of 2018 - 2019

In this article, we introduce data mining application to analyze enrollment data based on academic records at Phu Yen University (ĐHPY). We will collect the enrollment data for the school year of 2018 - 2019 at ĐHPY, and then carry out the step of building the database, data preprocessing and bringing the data to the table structure. We propose using Decision Tree algorithm of learning from the database to extract important rule sets relating to the enrollment. The result obtained after the extract can provide useful information for college-university admissions and organize the enrollment plan for University entrance exam year.

Keyword: Data mining, Decision Tree algorithm, University entrance exam

1. Giới thiệu

Năm học 2018 – 2019 là năm áp dụng cả 2 phương pháp xét tuyển cao đẳng đại học là xét điểm thi trung học phổ thông và xét điểm trung bình lớp 12. Đồng thời, các thí sinh đăng ký nhiều nguyện vọng học tập cho nhiều trường, chính vì lý do này làm ảnh hưởng rất lớn đến việc tuyển sinh ở các trường cao đẳng, đại học trên toàn quốc.

Với những lý do trên, chúng tôi đề xuất phương pháp phát hiện các yếu tố ảnh hưởng đến việc quyết định nhập học tại Trường Đại học Phú Yên (ĐHPY) dựa vào việc ứng dụng công nghệ khai phá dữ liệu (data mining). Các bước thực hiện nghiên cứu của chúng tôi bao gồm thu thập dữ liệu tuyển sinh cao đẳng đại học xét điểm học bạ, sau đó thực hiện xây dựng cơ sở dữ liệu, tiền xử lý dữ liệu, đưa dữ liệu ra cấu trúc bảng và dùng giải thuật cây quyết định để huấn luyện, rút trích dữ liệu các yếu tố ảnh hưởng đến việc nhập học. Kết quả

* Email: duyen.th2@gmail.com

thu được sau khi rút trích liên quan đến: học lực, khu vực, năm sinh, ngành học.

Nghiên cứu ứng dụng khai phá dữ liệu vào quản lý giáo dục đào tạo được xem rất cần thiết cho các nhà quản lý giáo dục, giúp công tác quản lý và hoạch định chiến lược giáo dục ngày càng hiệu quả. Gần đây có các công trình nghiên cứu ứng dụng khai phá dữ liệu trong giáo dục.

Nghiên cứu của (Nhuong, 2012) [1] đề xuất sử dụng khai phá dữ liệu về kết quả học tập của học sinh trường Cao đẳng nghề Văn Lang Hà Nội. Nghiên cứu một số kỹ thuật phân cụm trong khai phá dữ liệu: phân cụm phân chia, phân cụm dựa trên mật độ và phân cụm dựa trên lưới. Kết quả đạt được khi tiến hành áp dụng các giải thuật khai phá dữ liệu để rút trích được các yếu tố ảnh hưởng đến kết quả học tập của người học tại trường Cao đẳng nghề Văn Lang Hà Nội.

Nghiên cứu của (Nghị, 2014) [2] đề xuất sử dụng giải thuật rừng ngẫu nhiên học từ dữ liệu để rút trích các môn học quan trọng trong chương trình đào tạo ngành Công nghệ thông tin. Kết quả thu được sau khi rút trích có thể cung cấp thông tin hữu ích cho các nhà quản lý giáo dục trong việc tổ chức giảng dạy để nâng cao hiệu quả đào tạo.

Nghiên cứu của (Vinh, 2014) [3] đề xuất ứng dụng khai phá dữ liệu chọn ngành nghề cho học sinh THPT. Đề tài này sử dụng thuật toán ID3, cho ra mô hình phân lớp là một tập luật dưới dạng cây rất đơn giản và dễ hiểu, có độ chính xác khá cao và thời gian chạy chấp nhận được. Thuật toán được xây dựng trên nền website, thuận tiện cho người sử dụng ở bất cứ nơi đâu có kết nối Internet.

Các nghiên cứu trên đây đã tập trung vào việc dự đoán kết quả học tập, định hướng chọn ngành nghề. Nghiên cứu của chúng tôi đề xuất không đi theo hướng dự đoán chính xác kết quả, mà chúng tôi quan tâm đến việc phát hiện các yếu tố ảnh hưởng đến kết quả nhập học dựa trên giải thuật cây quyết định.

Phần tiếp theo của bài viết này được trình bày như sau: Phần 2 trình bày giải thuật cây quyết định; Phần 3 trình bày các kết quả thực nghiệm; Phần 4 trình bày kết luận và hướng phát triển.

2. Giải thuật cây quyết định

Cuối những năm 70 đầu những năm 80, J.Ross Quinlan [4] đã xây dựng một thuật toán sinh cây quyết định. Đây là một tiếp cận tham lam, trong đó nó xác định một cây quyết định được xây dựng từ trên xuống một cách đệ quy theo hướng chia để trị. Hầu hết các thuật toán sinh cây quyết định đều dựa trên tiếp cận top-down trình bày sau đây, trong đó nó bắt đầu từ một tập các bộ huấn luyện và các nhãn phân lớp của chúng. Tập huấn luyện được chia nhỏ một cách đệ quy thành các tập con trong quá trình cây được xây dựng.

Generate_decision_tree: Thuật toán sinh cây quyết định từ các bộ dữ liệu huấn luyện của nguồn dữ liệu D

Đầu vào:

- Nguồn dữ liệu D, trong đó có chứa các bộ dữ liệu huấn luyện và các nhãn phân lớp
- Attribute_list - danh sách các thuộc tính
- Attribute_selection_method, một thủ tục để xác định tiêu chí phân chia các bộ dữ liệu một cách tốt nhất thành các lớp. Tiêu chí này bao gồm một thuộc tính phân chia

splitting_attribute, điểm chia split_point và tập phân chia splitting_subset.

Đầu ra: Một cây quyết định

Nội dung thuật toán:

1. Tạo nút N
2. **If** các bộ trong D đều có nhãn lớp C **then**
3. Trả về N thành một nút lá với nhãn lớp C
4. **If** danh sách thuộc tính attribute_list là rỗng **then**
5. Trả về N thành một nút lá với nhãn là lớp chiếm đa số trong D (Việc này thực hiện qua gọi hàm **Attribute_selection_method**(D, attribute_list) để tìm ra tiêu chí phân chia tốt nhất splitting_criterion và gán nhãn cho N tiêu chí đó)
6. **If** splitting_attribute là một giá trị rời rạc và có nhiều cách chia **then**
7. Attribute_list = attribute_list – splitting_attribute // Loại bỏ thuộc tính splitting_attribute
8. **Foreach** j insplitting_criterion
9. // Phân chia các bộ xây dựng cây cho các phân chia đó
10. Đặt D_j là tập các bộ trong D phù hợp với tiêu chí j
11. **If** D_j là rỗng **then**
12. Gán nhãn cho nút N với nhãn phổ biến trong D
13. **Else** Gán nút được trả về bởi hàm Generate_decision_tree(D_j , attribute_list) cho nút N
14. **Endfor**
15. **Return** N

Lựa chọn thuộc tính

Việc lựa chọn thuộc tính sẽ phân tách tốt nhất các mẫu trong tập dữ liệu D vào các lớp phân biệt. Nếu chúng ta chia D thành các vùng nhỏ hơn dựa trên các kết quả tìm được của tiêu chí phân chia, thì mỗi vùng sẽ khá là thuần chủng (Nghĩa là các tập các vùng đã phân chia có thể hoàn toàn thuộc về cùng một lớp). Điều này giúp xác định cách các bộ giá trị tại một nút xác định sẽ được chia thế nào. Cây được tạo cho phân vùng D được gán nhãn với tiêu chí phân chia, các nhánh của nó được hình thành căn cứ vào các kết quả phân chia của các bộ.

Giả sử D là một phân vùng dữ liệu chứa các bộ huấn luyện được gán nhãn. Các nhãn có m giá trị phân biệt xác định m lớp, C_i (với $i = 1, \dots, m$). Gọi $C_{i,D}$ là tập các bộ của lớp C_i trong D

Thông tin cần thiết để phân lớp một bộ trong D cho bởi

$$\text{Info}(D) = - \sum_{i=1}^N p_i \log_2(p_i)$$

Trong đó p_i là khả năng một bộ trong D thuộc về lớp C_i được xác định bởi $|C_{i,D}|/|D|$.

Giờ giả sử chúng ta phân chia các bộ D dựa trên một số thuộc tính A có v giá trị phân biệt $\{a_1, \dots, a_v\}$. Thuộc tính A có thể dùng để chia D thành v phân vùng hoặc tập con $\{D_1, D_2, \dots, D_v\}$ trong đó D_j chứa các bộ trong D có kết quả đầu ra a_j . Các phân vùng đó sẽ tương đương với các nhánh của nút N.

Thông tin xác định xem việc phân chia đã gần tiếp cận đến một phân lớp được cho như sau

$$\text{Info}_A(D) = - \sum_{j=1}^N \frac{|D_j|}{|D|} \times \text{Info}(D_j)$$

$\frac{|D_j|}{|D|}$ là trọng lượng của phân vùng thứ j. $\text{Info}_A(D)$ thể hiện thông tin cần thiết để phân lớp một bộ của D dựa trên phân lớp theo A. Giá trị thông tin nhỏ nhất sẽ cho ra phân vùng thuần túy tương ứng.

Độ đo thông tin thu được được cho

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

$\text{Gain}(A)$ sẽ cho chúng ta biết bao nhiêu nhánh có thể thu nhận được từ A. Thuộc tính A với độ đo thông tin thu được lớn nhất sẽ được dùng làm thuộc tính phân chia của nút N.

3. Kết quả thực nghiệm

Trong phần thực nghiệm, chúng tôi tiến hành thu thập dữ liệu tuyển sinh năm học 2018 -2019 tại Trường Đại học Phú Yên theo hình thức xét điểm học bạ. Dữ liệu thu thập bao gồm danh sách đăng ký xét học bạ hệ cao đẳng và hệ đại học; danh sách sinh viên trúng tuyển nhập học năm học 2018 – 2019. Dữ liệu thu thập được có cấu trúc bảng được lưu trữ dưới dạng các file Microsoft Excel, dữ liệu này được thu thập vào ngày 25 tháng 9 năm 2018 (nguồn dữ liệu: Phòng Đào Tạo và Phòng Công Tác Học Sinh Viên – Trường Đại học Phú Yên).

Mỗi tập tin “Danh sách đăng ký xét học bạ Đại học Cao đẳng” chứa các thông tin: Mã học sinh, Họ và tên, giới tính, ngày sinh, khu vực, đối tượng, điểm trung bình lớp 12, xếp loại học lực, ...

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Danh sách đăng ký xét học bạ Đại học Cao đẳng															
2																
3	Mã	Họ và tên	GT	Ngày sinh	KV	ĐT	CMND	Mã sinh	Mã tuyển	Mã sinh	Mã trường	ĐTB km.12	Xếp loại học lực	Nguyễn		
4														Ngành		
5	1	Phạm Thị Thanh Xuân	1	04/5/2000	2		221495986	03	39	012	7.1	Khá	Giáo dục mầm non			
6	2	Nguyễn Quỳnh Vi	1	28/7/2000	2NT		221506116	39	07	39	020	7.9	Khá	Giáo dục mầm non		
7	3	Phạm Nguyễn Yến Như	1	06/8/2000	2		221476243	39	01	39	017	7.7	Khá	Kế toán		
8	4	Trần Thị Bích Thư	1	25/11/2000	2		221478791	39	01	39	007	7.8	Khá	Giáo dục mầm non		
9	5	Nguyễn Gia Tú	0	21/4/1999	2		221489155	39	01	39	004	7.3	TB	Quản trị kinh doanh		
10	6	Đỗ Thị Tường Vi	1	08/12/2000	2		221478788	39	07	39	007	7.4	Khá	Giáo dục mầm non		
11	7	Trần Lê Khả Hân	1	27/6/2000	2		221478783	39	01	39	007	7.1	Khá	Giáo dục mầm non		
12	8	Phan Thị Hoa	1	20/3/1999	2		221478024	39	01	39	007	7.6	Khá	Giáo dục mầm non		
13	9	Phạm Thị Nga	1	28/02/2000	2NT		221499198	39	09	39	021	8.2	Khá	Giáo dục mầm non		
14	10	Ngô Thị Ánh Nguyệt	1	04/12/2000	2NT		221491192	39	09	39	008	8.0	Giỏi			
15	11	Phan Huỳnh Thảo Ly	1	06/12/2000	2		221479193	39	01	39	017	7.0	Khá	Quản trị kinh doanh		
16	12	Tạ Thị Minh Thùy	1	15/9/1999	2NT		221482351	39	09	39	008	8.0	Khá	Giáo dục mầm non		
17	13	Huỳnh Văn Sang	0	02/6/1997	2NT		221448086	39	08	39	002	7.3	Khá	Chăn nuôi		

Hình 1. Cấu trúc tập tin Danh sách đăng ký xét học bạ Đại học Cao đẳng

Tập tin “Danh sách sinh viên trúng tuyển nhập học năm học 2018 – 2019” chứa các thông tin: ngành đào tạo, số báo danh, họ và tên, ngày sinh, giới tính, nơi sinh, hộ khẩu thường trú, đối tượng, khu vực, tổng điểm, đợt, mã số sinh viên.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	UBND TỈNH PHÚ YÊN				CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM										
2	TRƯỜNG ĐẠI HỌC PHÚ YÊN				Độc lập - Tự do - Hạnh phúc										
3															
4	DANH SÁCH THÍ SINH TRÚNG TUYỂN NHẬP HỌC TRÌNH ĐỘ ĐẠI HỌC														
5	Ngành: Giáo dục Mầm non (7140201)					Năm học: 2018 - 2019					Khóa học: 2018 - 2022				
6															
7	TT	SBD	HỌ VÀ TÊN		NGÀY SINH	GIỚI TÍNH	NƠI SINH	HỌ KHẨU THƯỜNG TRÚ	ĐỐI TƯỢNG		KHU VỰC	TỔNG ĐIỂM	NĂM TN THPT	ĐỢT	MÃ SỐ SINH VIÊN
8									ĐT	THUỘC DIỆN					
9	1		Nguyễn Thị Thu	Hiền	24/11/1999	Nữ	Phú Yên	Hòa Trị, Phú Hòa, Phú Yên			2NT	8.10	2017	1HB	18571402010001
10	2		Phan Ngọc	Hồng	10/05/2000	Nữ	Phú Yên	Sơn Phước, Sơn Hòa, Phú Yên			1	8.80	2018	1HB	18571402010002
11	3		Nguyễn Thị	Hợp	20/06/1999	Nữ	Phú Yên	Hòa Phú, Tây Hòa, Phú Yên			2NT	8.40	2017	1HB	18571402010003
12	4		Nguyễn Thị Thu	My	24/04/2000	Nữ	Phú Yên	Hòa Trị, Phú Hòa, Phú Yên			2NT	8.30	2018	1HB	18571402010005
13	5		Ngô Thị Ánh	Nguyệt	04/12/2000	Nữ	Phú Yên	Hòa Bình 1, Tây Hòa, Phú Yên			2NT	8.00	2018	1HB	18571402010007
14	6		Trần Thị Xuân	Phần	24/09/2000	Nữ	Phú Yên	Hòa Trị, Phú Hòa, Phú Yên			2NT	8.60	2018	1HB	18571402010009
15	7		Trần Thị Mỹ	Quyên	05/04/2000	Nữ	Phú Yên	Phường 9, TP.Tuy Hòa, Phú Yên			2	8.20	2018	1HB	18571402010010
16	8		Phan Thị	Thắm	20/03/2000	Nữ	Phú Yên	Phường 9, TP.Tuy Hòa, Phú Yên			2	8.40	2018	1HB	18571402010012
17	9		Nguyễn Phương	Thảo	10/02/2000	Nữ	Phú Yên	Xuân Quang 2, Đồng Xuân, Phú Yên			1	8.60	2018	1HB	18571402010013
18	10		Lê Thị Phương	Thu	16/03/2000	Nữ	Phú Yên	Phường 9, TP.Tuy Hòa, Phú Yên			2	8.10	2018	1HB	18571402010014
19	11	39006698	Nguyễn Thị Lam	Tường	27/07/2000	Nữ	Phú Yên	Hòa Trị, Phú Hòa, Phú Yên			2NT	23.15	2018	1THPT	18571402010018
20	12	39006243	Nguyễn Thị Thanh	Nhường	20/05/2000	Nữ	Phú Yên	Hòa Thắng, Phú Hòa, Phú Yên			2NT	22.27	2018	1THPT	18571402010019
21	13	39002914	Ksor Hò	Dim	20/10/2000	Nữ	Phú Yên	Ea Bía, Sông Hinh, Phú Yên	01	DT Đẻ	1	21.32	2018	1THPT	18571402010021

Hình 2. Cấu trúc tập tin Danh sách sinh viên trúng tuyển nhập học năm học 2018 – 2019

3.1. Xây dựng cơ sở dữ liệu

Dữ liệu thu thập được sẽ tổng hợp và chuyển dữ liệu từ file Microsoft Excel sang tổ chức lưu trữ dữ liệu của hệ quản trị cơ sở dữ liệu Microsoft SQL Server 2008. Chúng tôi xây dựng cơ sở dữ liệu tên là QLTuyenSinh dựa vào dữ liệu thu thập được. Cấu trúc cơ sở dữ liệu QLTuyenSinh gồm các bảng: *thisinh* (lưu danh sách đăng ký xét học bạ Cao đẳng Đại học), *nhaphoc* (lưu danh sách sinh viên trúng tuyển nhập học năm học 2018-2019) và *nganh* (lưu thông tin các ngành đào tạo).

Tập tin *thisinh* chứa các thông tin: cao đẳng đại học, Họ, Tên, Giới tính, Ngày sinh, Nơi sinh, Khu vực, Đối tượng, điểm trung bình, học lực, mã ngành.

Tập tin *nhaphoc* chứa các thông tin: mã ngành, cao đẳng đại học, Họ, Tên, ngày sinh, Giới tính, Ngày sinh, Nơi sinh, Khu vực, Đối tượng, điểm trung bình, học lực, mã sinh viên.

Tập tin *nganh* chứa các thông tin: mã ngành, tên ngành.

3.2. Tiền xử lý dữ liệu

Bước tiếp theo là tiền xử lý dữ liệu: dữ liệu sau khi chuyển sang hệ quản trị cơ sở dữ liệu sẽ được tổng hợp và chuyển về một bảng dữ liệu duy nhất, các cột gồm: mã ngành, mã thí sinh, giới tính, năm sinh, nơi sinh, khu vực, điểm trung bình, học lực, nhập học. Để làm được điều này chúng tôi tiến hành hai bước:

Bước 1: cập nhật lại dữ liệu nhập sai hoặc thiếu do quá trình nhập liệu từ file Microsoft Excel.

Bước 2: dựa trên dữ liệu vừa xây dựng ở bước 1, chúng tôi tiếp tục tạo bảng ảo (view) chứa các thông tin của danh sách xét học bạ và danh sách nhập học thành 1 bảng. Bảng ảo này có tên là xldltong (xử lý dữ liệu tổng) gồm các thông tin: Mã ngành, Giới tính, Năm sinh, Nơi sinh, Khu vực, điểm trung bình, học lực, nhập học.

Sau quá trình tiền xử lý dữ liệu, chúng tôi thu được bảng dữ liệu có 351 dòng và 8 trường.

manganh	GIOITINH	namsinh	NOISINH	KV	TONGDIEM	HOCLUC	NHAPHOC
6340404	Nam	1999	PHUYEN	KV2	73	KHA	HOC
6340404	Nu	1999	PHUYEN	KV2	67	TBKHA	HOC
6340404	Nu	1999	PHUYEN	KV2	67	KHA	KHONG
6340404	Nu	2000	PHUYEN	KV1	78	KHA	KHONG
6340404	Nu	2000	PHUYEN	KV2	80	GIOI	HOC
6340404	Nam	2000	PHUYEN	KV2NT	84	KHA	KHONG
6340404	Nu	1999	PHUYEN	KV1	59	TB	KHONG
6340404	Nam	1999	PHUYEN	KV2	73	TB	KHONG
6340404	Nu	1998	PHUYEN	KV2	64	TBKHA	HOC
6340404	Nu	1999	PHUYEN	KV2	72	KHA	HOC
6340404	Nu	1999	PHUYEN	KV2NT	72	KHA	KHONG

Hình 3. Cấu trúc tập tin sau khi tiền xử lý dữ liệu

Sau khi thực hiện tiền xử lý dữ liệu, chúng tôi được tập tin dữ liệu có cấu trúc bảng, sử dụng để phân tích dữ liệu tuyển sinh dựa vào xét học bạ tại Trường Đại học Phú Yên trong năm học 2018 – 2019.

3.3. Xây dựng mô hình cây quyết định

Chương trình chúng tôi dùng phần mềm WEKA 3.8.3 (New Zealand, 1999-2018) [5]. Tiến hành xây dựng mô hình cây quyết định theo thuật toán C4.5 được chương trình hóa trong mục phân lớp trên công cụ Weka là J48. Kết quả đánh giá dữ liệu thu được từ Weka là: trường hợp phân loại chính xác là 75,2% và trường hợp phân loại không chính xác là 24,8% như hình 4. Kết quả của quá trình xây dựng mô hình cây quyết định này như hình 5.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48 - C 0.25 - M 2**

Test options

☐ Use training set
☒ Supplied test set **Set...**
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

Classifier output

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.06 seconds

=== Summary ===

Correctly Classified Instances      264      75.2137 %
Incorrectly Classified Instances    87      24.7863 %
Kappa statistic                    0.5038
Mean absolute error                 0.3519
Root mean squared error             0.4194
Relative absolute error             70.3789 %
Root relative squared error         83.8921 %
Total Number of Instances          351

=== Detailed Accuracy By Class ===

```

(Nom) NHAPHOC

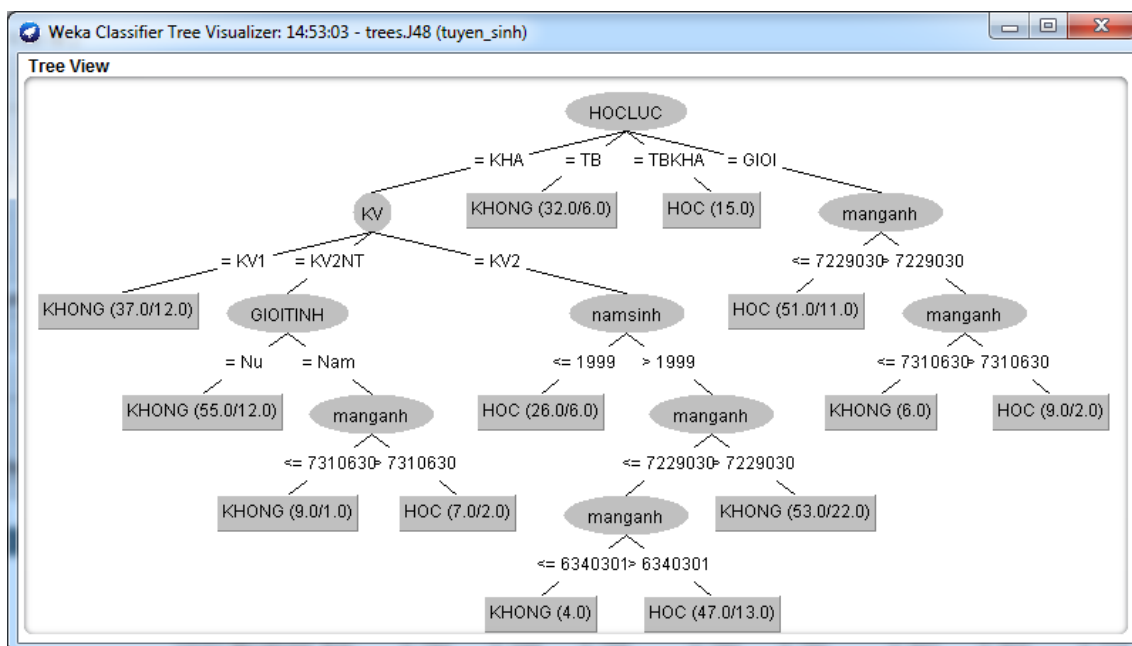
Start Stop

Result list (right-click for options)

14:27:04 - trees.J48
14:27:20 - trees.J48

TP Rate FP Rate Precision Recall F-Measure MCC

Hình 4. Kết quả đánh giá dữ liệu



Hình 5. Mô hình cây quyết định của dữ liệu tuyển sinh

Mô hình cây nhị phân này chúng ta có thể thấy được là học lực của thí sinh quyết định nhiều đến việc xác nhận thực học tại trường:

- ✓ Nếu học lực là trung bình thì không học tại trường.
- ✓ Nếu học lực là trung bình khá thì xác nhận nhập học tại trường.
- ✓ Nếu học lực là khá thì phải xét đến yếu tố khu vực của nhóm thí sinh có nhập học hay là không nhập học. Học lực là Khá và thí sinh là khu vực 1 thì không nhập học; nếu là khu vực 2 thì phải xét thêm yếu tố là năm sinh và ngành học; nếu là khu vực 2NT thì phải xét thêm yếu tố là giới tính là ngành học.
- ✓ Nếu học lực là giỏi thì xét đến yếu tố là ngành học.

Kết quả thực tế nhập học của sinh viên năm nhất tại trường tính đến ngày 25/09/2018, chúng tôi tiến hành kiểm tra tỷ lệ phần trăm số lượng sinh viên nhập học theo nhóm học lực so với tổng số sinh viên nhập học trên tập dữ liệu đã thu thập thì kết quả thực tế so với kết quả từ mô hình cây quyết định xây dựng theo giải thuật J48 như bảng 1.

Học lực	Tỷ lệ % sinh viên nhập học thực tế	Tỷ lệ % sinh viên nhập học theo mô hình đã xây dựng
Giỏi	27.01%	38.71%
Khá	60.92%	51.61%
Trung bình – Khá	3.45%	0%
Trung bình	8.62%	9.68%

Bảng 1. Thống kê nhập học thực tế của sinh viên năm nhất xét điểm học bạ

Theo như bảng 1 kết quả tương đối tương đồng giữa kết quả thực tiễn và kết quả rút

trích ra từ mô hình cây quyết định đã xây dựng. Do vậy, những tập luật rút trích được là những vấn đề đáng quan tâm đối với việc tuyển sinh.

4. Kết luận và hướng phát triển

Chúng tôi vừa trình bày một cách tiếp cận khai phá dữ liệu để phân tích dữ liệu tuyển sinh dựa vào xét học bạ tại trường Đại học Phú Yên trong năm học 2018 – 2019. Các bước thực hiện bao gồm thu thập dữ liệu tuyển sinh tại trường, tiếp theo là xây dựng cơ sở dữ liệu, tiền xử lý dữ liệu và xây dựng mô hình cây quyết định cho phép rút trích ra những tập luật quan trọng. Kết quả thu được sau khi rút trích các tập luật quan trọng, có thể cung cấp thông tin hữu ích cho việc tuyển sinh năm học 2019 -2020: dự đoán được đối tượng học sinh có thể quyết định học tại trường và nhà trường có thể lên kế hoạch tuyển sinh hợp lý.

Trong tương lai chúng tôi dự định mở rộng nghiên cứu và phát triển cho việc đánh giá kết quả học tập của sinh viên. Ngoài ra, cần phải tham khảo thêm nhiều ý kiến khác của các chuyên gia để góp phần nâng cao độ tin cậy trong việc tìm ra những tập luật quan trọng□

TÀI LIỆU THAM KHẢO

- [1] Nguyễn Đăng Nhượng (2012), *Khai phá dữ liệu về kết quả học tập của học sinh trường Cao đẳng nghề Văn Lang Hà Nội*, Thạc sĩ, Trường Đại học Công nghệ.
- [2] Đỗ Thanh Nghị, Phạm Nguyên Khang, Nguyễn Minh Trung và Trịnh Trung Hưng (2014), “Phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập sinh viên ngành Công Nghệ Thông Tin”, *Tạp chí Khoa học Trường Đại học Cần Thơ*, Số 33 (2014), Trang: 49-57.
- [3] Nguyễn Đặng Thế Vinh (2014), *Ứng dụng khai phá dữ liệu chọn ngành nghề cho học sinh THPT*, Thạc sĩ, Trường Đại học Quốc tế Hồng Bàng.
- [4] Jiawei Han, Micheline Kamber, Jian Pei (2012), *Data Mining: Concepts and Techniques, 3rd Edition*, Morgan Kaufmann.
- [5] <https://www.cs.waikato.ac.nz/ml/weka/downloading.html>

(Ngày nhận bài: 09/10/2018; ngày phản biện: 26/11/2018; ngày nhận đăng: 04/01/2019)