Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

PHÁT HIỆN MÔN HỌC QUAN TRỌNG ẢNH HƯỞNG ĐẾN KẾT QUẢ HỌC TẬP SINH VIÊN NGÀNH CÔNG NGHỆ THÔNG TIN

Đỗ Thanh Nghị[1], Phạm Nguyên Khang[1], Nguyễn Minh Trung[2] và Trịnh Trung Hưng[3]

[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
[2] *Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[3] *Trung tâm Liên kết Đào tạo, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 20/01/2014
Ngày chấp nhận: 28/08/2014

Title:
Detection of the key courses affecting the learning outcomes of information technology students

Từ khóa:
Chương trình đào tạo ngành CNTT, Khai mỏ dữ liệu, Rừng ngẫu nhiên, Rút trích đặc trưng

Keywords:
Study program of information technology, Data mining, Random forests, Feature extraction

ABSTRACT

This paper presents data mining approach for detecting the key courses which affect the learning outcomes of information technology students. We collect the study results of undergraduate students studying information technology programs at Can Tho University; and then the pre-processing step is to transform the dataset into structured one (i.e. the table format) suited for the input of data mining algorithms used in the next step. The random forest model is learnt from the dataset to extract the important features (the key courses). The experimental results showed that the key courses extracted by our proposed approach provide useful information to educational managers to improve the training efficiency.

TÓM TẮT

Trong bài này, chúng tôi giới thiệu tiếp cận khai mỏ dữ liệu để phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của sinh viên ngành công nghệ thông tin (CNTT). Chúng tôi tiến hành sưu tập dữ liệu học tập của sinh viên tốt nghiệp ngành CNTT tại Trường Đại học Cần Thơ, sau đó thực hiện bước tiền xử lý dữ liệu, đưa dữ liệu về cấu trúc bảng. Chúng tôi 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 CNTT. 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.

1 GIỚI THIỆU

Trong nhiều năm qua, ngay cả khi số lượng đào tạo nhân lực tại các trường đại học, cao đẳng chuyên ngành về công nghệ thông tin (CNTT) đã tăng gấp 3 cho đến 4 lần, nhu cầu nguồn nhân lực CNTT tăng nhanh. Nhưng theo đánh giá của các nhà tuyển dụng, đào tạo CNTT ở các trường hiện chưa đáp ứng được nhu cầu thực tiễn. Nguyên nhân chủ yếu do chất lượng đầu ra của sinh viên ngành CNTT vẫn còn thấp. Để nâng cao được chất lượng của sinh viên nhằm đáp ứng được nhu cầu thực tiễn, cần phải có sự phối hợp nhịp nhàng giữa nhà tuyển dụng, cơ sở đào tạo mà ở đó vai trò của nhà quản lý giáo dục, đội ngũ giảng viên, giáo viên cố vấn học tập và sinh viên. Làm sao giáo viên cố vấn học tập tư vấn để sinh viên biết được kiến thức nào là quan trọng ảnh hưởng đến kết quả khi ra trường. Nhờ đó sinh viên chú tâm hơn ở các môn học quan trọng nhằm cải thiện được chất lượng học tập. Đồng thời, nhà quản lý cũng có cơ hội bố trí, sắp xếp chương trình, đội ngũ giảng viên phù hợp với các môn học thuộc phần kiến thức quan trọng.

Chúng tôi đề xuất tiếp cận phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của sinh

viên CNTT tại Trường Đại học Cần Thơ (ĐHCT), dựa trên công nghệ khám phá tri thức và khai mỏ dữ liệu (Fayyad *et al*., 1996). Qua đó, nhà quản lý có chiến lược quản lý phù hợp nhằm cải tiến chất lượng giảng dạy cho nhóm môn học quan trọng, giáo viên cố vấn tư vấn cho sinh viên tập trung cải thiện chất lượng học tập. Nâng cao hiệu quả đầu ra của sinh viên CNTT. Các bước thực hiện nghiên cứu của chúng tôi bao gồm sưu tập dữ liệu học tập của sinh viên tốt nghiệp ngành CNTT, sau đó thực hiện bước tiền xử lý dữ liệu, đưa dữ liệu về cấu trúc bảng mà từ đó giải thuật rừng ngẫu nhiên (Breiman, 2001) được huấn luyện để rút trích các môn học quan trọng trong chương trình đào tạo. Kết quả thu được sau khi rút trích bao gồm các môn học như xác suất thống kê, toán rời rạc, cấu trúc dữ liệu, có thể cung cấp thông tin hữu ích cho các nhà quản lý giáo dục, giảng viên, sinh viên trong việc tổ chức giảng dạy để nâng cao hiệu quả đào tạo.

Phần tiếp theo của bài viết được trình bày như sau: Phần 2 trình bày ngắn gọn về các nghiên cứu liên quan; Phần 3 trình bày giải thuật học rừng ngẫu nhiên và các rút trích đặc trưng; Phần 4 trình bày các kết quả thực nghiệm tiếp theo sau đó là kết luận và hướng phát triển.

2 NGHIÊN CỨU LIÊN QUAN

Nghiên cứu ứng dụng khai mỏ 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 kỹ thuật khai mỏ dữ diệu đem lại nhiều lợi ích trong giáo dục.

Nghiên cứu của (Lê, 2002) đề xuất sử dụng khai phá luật kết hợp (Agrawal *et al*., 1993) và logic mờ (Zadeh, 1965) trên kết quả thi tốt nghiệp THPT và THCS cho mục tiêu đánh giá hiệu quả đào tạo và cung cấp các thông tin cần thiết cho quá trình nâng cao chất lượng học sinh. Một hướng tiếp cận tương tự của tác giả (Nguyễn, 2002) sử dụng luật kết hợp trong việc tính điểm để phát hiện học sinh yếu, các học sinh cần phụ đạo thêm.

Luận văn thạc sĩ của (Phan, 2009) đã nghiên cứu phương pháp khai mỏ tìm luật kết hợp trên dữ liệu giáo dục. Ứng dụng thực nghiệm trên dữ liệu kết quả học tập của sinh viên trường Đại học Tôn Đức Thắng, nhằm hỗ trợ đánh giá và dự đoán kết quả học tập của sinh viên, qua đó nâng cao chất lượng đào tạo.

Đề tài thạc sĩ của (Nguyễn, 2011) tập trung xây dựng hệ thống dự đoán tốt nghiệp phổ thông trung học. Tác giả áp dụng thuật toán khai phá luật kết hợp mờ vào việc dự đoán kết quả tốt nghiệp phổ thông trung học dựa trên học lực và hạnh kiểm của học sinh.

Nghiên cứu khác của (Nguyễn, 2012) trình bày kết quả đã đạt được khi tiến hành áp dụng giải thuật gom cụm dữ liệu, kMeans (MacQueen, 1967) để khai thác thông tin từ điểm học sinh của trường Cao đẳng nghề Văn Lang Hà Nội. Tác giả tìm hiểu sự ảnh hưởng của vùng miền, của hoàn cảnh gia đình, dân tộc, đạo đức… đến kết quả học tập của học sinh, phân loại kết quả học tập để đánh giá một cách nhanh chóng nhận thức của người học. Từ đó có những điều chỉnh giảng dạy của giáo viên phù hợp với năng lực người học.

Nghiên cứu của (Nguyễn *et al*., 2007) đề xuất sử dụng giải thuật máy học cây quyết định (Breiman *et al*., 1984), (Quinlan, 1993) và mạng Bayes (Pearl, 1985) trong dự đoán kết quả học tập của sinh viên đại học và sau đại học của Trường ĐHCT. Một nghiên cứu khác của (Nguyễn *et al*., 2011) đề xuất sử dụng kỹ thuật phân rã ma trận để dự đoán kết quả học tập của sinh viên.

Nghiên cứu của (Pal & Pal, 2013) đề xuất sử dụng giải thuật máy học cây quyết định (Breiman *et al*., 1984), (Quinlan, 1993) và Bagging (Breiman, 1996) để dự đoán kết quả học tập của sinh viên Đại học Purvanchal, Ấn Độ.

Nghiên cứu của (Bukralia *et al*., 2012) đã đề xuất sử dụng các kỹ thuật máy học như mạng nơ-ron, hồi quy logistic (Hastie *et al*., 2001), cây quyết định (Breiman *et al*., 1984), (Quinlan, 1993), máy học véc-tơ hỗ trợ SVM (Vapnik, 1995) để dự đoán kết quả học tập của sinh viên theo hệ đào tạo từ xa của Đại học Midwest, Hoa Kỳ.

Có thể thấy được rằng, các nghiên cứu trên đây đều tập trung vào dự đoán kết quả học tập, dự đoán điểm môn học. Nghiên cứu của chúng tôi đề xuất không đi theo hướng dự đoán chính xác kết quả học tập. Chúng tôi quan tâm đến phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của sinh viên ngành CNTT dựa trên giải thuật học rừng ngẫu nhiên.

3 GIẢI THUẬT RỪNG NGẪU NHIÊN

Tiếp cận rừng ngẫu nhiên do (Breiman, 2001) đưa ra là một trong những phương pháp tập hợp mô hình thành công nhất. Giải thuật rừng ngẫu nhiên tạo ra một tập hợp các cây quyết định

(Breiman *et al*., 1984), (Quinlan, 1993) không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap (lấy mẫu có hoàn lại từ tập học), tại mỗi nút phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính. Lỗi tổng quát của rừng phụ thuộc vào độ chính xác của từng cây thành viên trong rừng và sự phụ thuộc lẫn nhau giữa các cây thành viên. Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho thành phần lỗi bias thấp (thành phần lỗi bias là thành phần lỗi của giải thuật học, nó độc lập với tập dữ liệu học) và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng. Tiếp cận rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay. Như trình bày trong (Breiman, 2001), rừng ngẫu nhiên học nhanh, chịu đựng nhiễu tốt và không bị tình trạng học vẹt. Giải thuật rừng ngẫu nhiên sinh ra mô hình có độ chính xác cao đáp ứng được yêu cầu thực tiễn cho vấn đề phân loại, hồi qui.

3.1 Giải thuật xây dựng rừng ngẫu nhiên

Giải thuật máy học rừng ngẫu nhiên (Hình 1) có thể được trình bày ngắn gọn như sau:

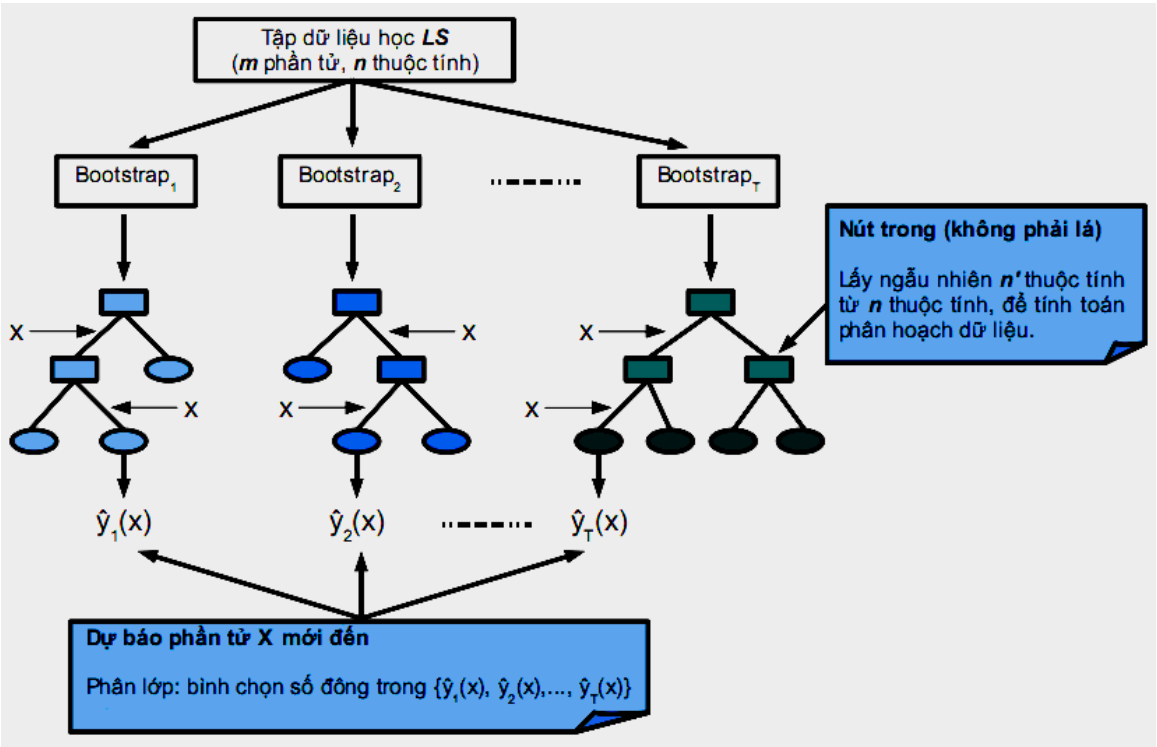
– Từ tập dữ liệu học ***LS*** có ***m*** phần tử và ***n*** biến (thuộc tính), xây dựng ***T*** cây quyết định một cách độc lập nhau

– Mô hình cây quyết định thứ t được xây dựng trên tập mẫu Bootstrap thứ ***t*** (lấy mẫu ***m*** phần tử có hoàn lại từ tập học ***LS***)

– Tại nút trong, chọn ngẫu nhiên ***n'*** biến (***n'***<<***n***) và tính toán phân hoạch tốt nhất dựa trên ***n'*** biến này

– Cây được xây dựng đến độ sâu tối đa không cắt nhánh

Kết thúc quá trình xây dựng ***T*** mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử mới đến ***X***.



Hình 1: Giải thuật rừng ngẫu nhiên

3.2 Rút trích đặc trưng

Rút trích đặc trưng quan trọng được thực hiện trong khi huấn luyện mô hình của rừng ngẫu nhiên. Mỗi bước ***t***, sử dụng tập ***Bootstrap$_t$*** (lấy mẫu có hoàn lại ***m*** phần tử từ tập huấn luyện ***LS***) để xây dựng mô hình cây quyết định cơ sở thứ ***t*** (***DT$_t$***) trong rừng ngẫu nhiên; giải thuật lấy tập Out-Of-Bootstrap$_t$, ***OOB$_t$*** (các phần tử trong tập dữ liệu huấn luyện ***LS*** nhưng không nằm trong tập ***Bootstrap$_t$***) làm tập kiểm tra để tính độ chính xác phân lớp của cây ***DT$_t$*** trong rừng ngẫu nhiên.

Thuộc tính quan trọng được hiểu là thuộc tính làm ảnh hưởng rất nhiều đến kết quả phân lớp của rừng ngẫu nhiên. Cụ thể là nếu có những thay đổi (hoán vị các giá trị của thuộc tính) thì độ chính xác phân lớp của rừng ngẫu nhiên bị giảm nhiều so với khi chưa tác động làm thay đổi thuộc tính.

Việc thực hiện các tính toán để xác định thuộc tính quan trọng trong rừng ngẫu nhiên như sau. Khi xây dựng cây thứ ***t*** học từ tập ***Bootstrap$_t$***, tính độ chính xác của cây ***DT$_t$*** sử dụng tập ***OOB$_t$*** (Out-Of-Bootstrap), là ***Acc(DT$_t$, OOB$_t$)***. Lần lượt thực hiện hoán vị giá trị của từng thuộc tính thứ ***i*** của tập ***OOB$_t$***, là ***OOB$_t$(rand(i))***. Tính lại độ chính xác của cây ***DT$_t$*** sử dụng tập ***OOB$_t$(rand(i))***, ***Acc(DT$_t$, OOB$_t$ (rand(i))***. Tiếp đến, cần tính lại sự khác biệt về độ chính xác trước và sau khi hoán vị các giá trị của thuộc tính thứ ***i*** của cây ***DT$_t$***. Với các thuộc tính ***i=1, 2, …, k***, chúng ta có:

$\Delta acc_{t,1} = Acc(DT_t, OOB_t) - Acc(DT_t, OOB_t(rand(1)))$

$\Delta acc_{t,2} = Acc(DT_t, OOB_t) - Acc(DT_t, OOB_t(rand(2)))$

……

$\Delta acc_{t,k} = Acc(DT_t, OOB_t) - Acc(DT_t, OOB_t(rand(k)))$

Với mô hình rừng ngẫu nhiên ***RF*** có ***T*** cây, chúng ta có được tổng sự khác biệt về độ chính xác trước và sau khi hoán vị các giá trị của thuộc tính của rừng ngẫu nhiên ***RF*** là:

thuộc tính 1: $\alpha_1 = \Delta acc_{1,1} + \Delta acc_{2,1} + ... + \Delta acc_{T,1}$

thuộc tính 2: $\alpha_2 = \Delta acc_{1,2} + \Delta acc_{2,2} + ... + \Delta acc_{T,2}$

.....

thuộc tính k: $\alpha_k = \Delta acc_{1,k} + \Delta acc_{2,k} + ... + \Delta acc_{T,k}$

Sắp xếp $\alpha_1, \alpha_2, ..., \alpha_k$ theo thứ tự giảm dần, điều này đồng nghĩa với thứ tự tổng sự khác biệt về độ chính xác trước và sau khi hoán vị các giá trị của các thuộc tính. Sự khác biệt càng lớn thì thuộc tính tương ứng càng quan trọng. Từ ý tưởng này, chúng ta thực hiện rút trích môn học quan trọng ảnh hưởng đến kết quả học tập của sinh viên ngành CNTT. Chúng ta có thể xem sinh viên như là 1 dòng (mẫu tin, phần tử của dữ liệu), các môn học của sinh viên xem như thuộc tính (cột, trường) và kết quả xếp loại học tập khi ra trường có thể xem là lớp (nhãn). Như vậy, dữ liệu học tập của sinh viên chính là bảng dữ liệu. Chúng tôi sử dụng rừng ngẫu nhiên học để phân loại sinh viên. Trong quá trình xây dựng mô hình học, rừng ngẫu nhiên thực hiện rút trích các môn học (thuộc tính) quan trọng như vừa được mô tả. Có thể diễn giải rằng những môn học quan trọng được rút trích từ mô hình học rừng ngẫu nhiên là những môn học làm ảnh hưởng rất lớn đến kết quả phân loại học tập của sinh viên.

4 KẾT QUẢ THỰC NGHIỆM

Trong phần thực nghiệm, chúng tôi tiến hành thu thập dữ liệu kết quả học tập của sinh viên tại phòng đào tạo, Trường ĐHCT. Dữ liệu thu thập bao gồm kết quả học tập của sinh viên ngành CNTT thuộc các khóa từ 20 đến 29 (tuyển sinh từ năm 1994 đến 2003). Các khóa từ 30 trở về sau được điều chỉnh bởi quy chế đào tạo tín chỉ sử dụng phương pháp đánh giá kết quả học tập theo thang điểm chữ. (A, B, C,..) nên dữ liệu không đồng nhất, do số lượng dữ liệu thu thập đã đủ lớn để nghiên cứu nên chúng tôi không thu thập dữ liệu các khóa này. Dữ liệu thu thập được có dạng cấu trúc bảng, được tổ chức theo từng học kỳ năm học. Mỗi học kỳ năm học có các tập tin dữ liệu như: *diem* (lưu điểm sinh viên), *dtotng* (lưu sinh viên tốt nghiệp), *ctdt* (lưu chương trình đào tạo)…và các tập tin dữ liệu khác tại học kỳ đó. Bên cạnh đó, còn có các tập tin lưu trữ thông tin diễn giải các mã số của hệ thống như: họ tên sinh viên, tên môn học, tên ngành,…

Mỗi tập tin *diem* chứa các thông tin: mã số sinh viên, mã số môn học (tùy chọn và bắt buộc), điểm thi của các môn học,… mà sinh viên tham gia học vào từng học kỳ. (Hình 2)

Diem

F_masv	F_mamh	F_manh	F_diembt	F_to	F_diem1	F_diem2	F_diemtl
1980001	TN002C	04				4.5	
1980001	ML101C	04				6.0	
1980001	VL002C	04				7.0	
1980001	TH001C	04				7.0	
1980001	HH002C	04				3.0	
1980001	TN006C	04				8.0	
1980001	TH004C	04				5.0	
1980001	VL092C	04				6.0	
1980002	TN002C	04				5.0	
1980002	ML101C	04				6.5	
1980002	VL002C	04				4.0	
1980002	TH001C	04				6.0	
1980002	HH002C	04				4.5	
1980002	TN006C	04				6.0	
1980002	TH004C	04				3.0	
1980002	VL092C	04				6.0	
1980003	TN002C	04				4.0	
1980003	ML101C	04				4.5	
1980003	VL002C	04				5.0	
1980003	TH001C	04				5.0	
1980003	HH002C	04				4.0	
1980003	TN006C	04				5.0	
1980003	TH004C	04				5.0	
1980003	VL092C	04				5.0	

Hình 2: Cấu trúc tập tin điểm

Tập tin *dtotng* chứa thông tin kết quả xếp loại tốt nghiệp của sinh viên bao gồm: mã số sinh viên, mã ngành học, điểm tốt nghiệp, xếp loại tốt nghiệp,…(Hình 3)

Dtotng

F_masv	F_mang	F_mahedt	F_nhhk	F_dtbtn	F_xeploai	F_dottn
1950233	56			5.61	Trung Bình	30074
1960193	56			5.44	Trung Bình	30064
1960216	56			5.46	Trung Bình	30064
1970423	56			6.12	Tbình_ Khá	30084
1970483	56			5.97	Trung Bình	30074
1970513	56			6.24	Tbình_ Khá	30064
1970525	56			5.71	Trung Bình	30074
1970548	56			6.60	Tbình_ Khá	30064
1970554	56			6.07	Tbình_ Khá	30074
1980515	56			6.17	Tbình_ Khá	30074
1980521	56			6.33	Tbình_ Khá	30084
1980522	56			6.06	Tbình_ Khá	30084
1980551	56			6.03	Tbình_ Khá	30084
1980561	56			6.20	Tbình_ Khá	30074
1980581	56			6.04	Tbình_ Khá	30074
1980597	56			6.65	Tbình_ Khá	30084

Hình 3: Cấu trúc tập tin điểm tốt nghiệp

Ngoài ra, chúng tôi còn sử dụng tập tin *ctdmh* (lưu tên môn học) để diễn giải tên môn học từ mã số môn học (Hình 4)

Ctdmh

F_mamh	F_tenmhvn	F_dvht	F_ts	F_lt
TH343C	Đồ án môn học 2 - Tin học	2	30	0
TH344C	Xử lý tín hiệu số	3	45	45
TH345C	Cơ sở viễn thông	4	60	45
TH346C	TT.Kỹ thuật đo & vi xử lý	2	45	0
TH347C	Điện tử công suất & ứng dụng	2	30	30
TH348C	TT.Điện tử công suất &ứng dụng	1	30	0
TH349C	Kỹ thuật Audio & video	5	75	75
TH350C	An-ten & truyền sóng	3	45	45
TH351C	Đồ án môn học 1 - Điện tử	2	30	0

Hình 4: Cấu trúc tập tin môn học

Bên cạnh đó, chúng tôi cũng tìm hiểu phương pháp tính điểm học tập của sinh viên để xếp loại tốt nghiệp. Điểm tốt nghiệp (ĐTN): là trung bình có trọng số của điểm các môn học đã tích lũy tính đến thời điểm xét (không bao gồm các môn học điều kiện, và các môn học bị điểm F). Công thức như sau:

$$ĐTN = \frac{\sum_{i=1}^{n} a_i X_i}{\sum_{i=1}^{n} a_i} \quad (1)$$

Với X_i là điểm môn học thứ i; a_i là số tín chỉ của môn học thứ i; n là số môn học sinh viên tích lũy.

4.1 Tiền xử lý dữ liệu

Bước tiếp theo là tiền xử lý dữ liệu: dữ liệu thu thập được sẽ được tổng hợp và chuyển về một bảng dữ liệu duy nhất, mỗi cột (trường - field) của bảng biểu diễn tên của mỗi môn học, mỗi dòng (record) mô tả kết quả học tập toàn khóa học của mỗi sinh viên. Để làm được điều này chúng tôi tiến hành thực hiện hai bước:

Ten_MH	Diem	Diem_TB	Diem_TB_TN
Niên luận 1-Tin học (Lập trình	7	9.50	7.69
Niên luận 1-Tin học (Lập trình	5.5	9.50	7.4
Niên luận 2 - Tin học	9	9.50	7.4
Niên luận 2 - Tin học	8.5	9.50	7.69
Niên luận 3-Tin học (XD HTTT)	9	9.50	7.69
Niên luận 3-Tin học (XD HTTT)	8	9.50	7.4
Phân tích hệ thống	7.5	9.50	7.4
Phân tích hệ thống	8	9.50	7.69
Phân tích&thiết kế thuật toán	10	9.50	7.69
Phân tích&thiết kế thuật toán	9.5	9.50	7.4

Hình 5: Dữ liệu tổng hợp

– Bước 1: lọc lấy dữ liệu điểm của sinh viên ngành CNTT, xóa bỏ các sinh viên ngành khác; xóa bỏ các dữ liệu không hợp lệ; các thuộc tính không quan trọng và chuyển dữ liệu từ các tập tin điểm và tập tin tốt nghiệp về một bảng duy nhất (như bảng ở Hình 5). Ở bước này chúng ta thu được một bảng đã bỏ qua các thuộc tính không quan trọng ví dụ như mã số, họ tên, tên ngành, …

– Bước 2: Dựa trên bảng dữ liệu vừa xây dựng ở bước 1, chúng tôi tiếp tục biểu diễn dữ liệu trong bảng sao cho các cột mô tả các môn học, các dòng mô tả điểm các môn học mà mỗi sinh viên có thể tham gia trong suốt khóa học. Vì có một số môn học sinh viên không tham gia học (có thể các khóa có các môn học khác nhau) hoặc được miễn, trường hợp này điểm của các môn này có giá trị là NULL và chúng tôi xử lý các giá trị NULL thành giá trị 'NA'. Sau quá trình tiền xử lý dữ liệu, chúng tôi thu được bảng dữ liệu có 317 dòng và 249 trường (mỗi trường tương ứng với một môn học, giá trị của mỗi trường là điểm của môn học đó) và trường cuối cùng là điểm trung bình tốt nghiệp (như Hình 6).

An toà...	Anh văn ...	C.Đề 1: Xử lý...	C.Đề 2: Trí tuệ ...	Tin học ...	Toán rời rạc...	Trí tuệ ...	Vi-Tích ...	Xác suất th...	Xây dựng hệ thống...	Xử lý ảnh	Diem_TB	Diem_TB_TN
NA	5.5	NA	NA	5.5	NA	5	4.5	NA	5	NA	6.19	5.96
NA	8	NA	NA	4.5	3.5	5	4	NA	NA	NA	9.00	6.03
NA	6.5	NA	NA	4	NA	5	5.5	NA	5	0.5	6.37	6.02
NA	7	NA	NA	4	NA	5	7.5	NA	NA	0	7.50	6.25
NA	8	NA	NA	6	NA	0	9	NA	NA	7.5	6.71	6.48
NA	7	NA	NA	4	NA	0	4	NA	NA	NA	6.75	6.11
NA	7.5	NA	NA	4.5	NA	NA	4	NA	NA	NA	6.17	5.7
NA	8.5	NA	NA	6	NA	0	8	NA	NA	0	6.75	6.1
NA	7	NA	NA	3.5	NA	NA	8	NA	NA	1.5	7.85	6.12
NA	NA	NA	NA	NA	9	4	7.5	NA	NA	NA	9.33	7.28
NA	8	NA	NA	NA	8	5	4.5	NA	NA	NA	7.90	7.17

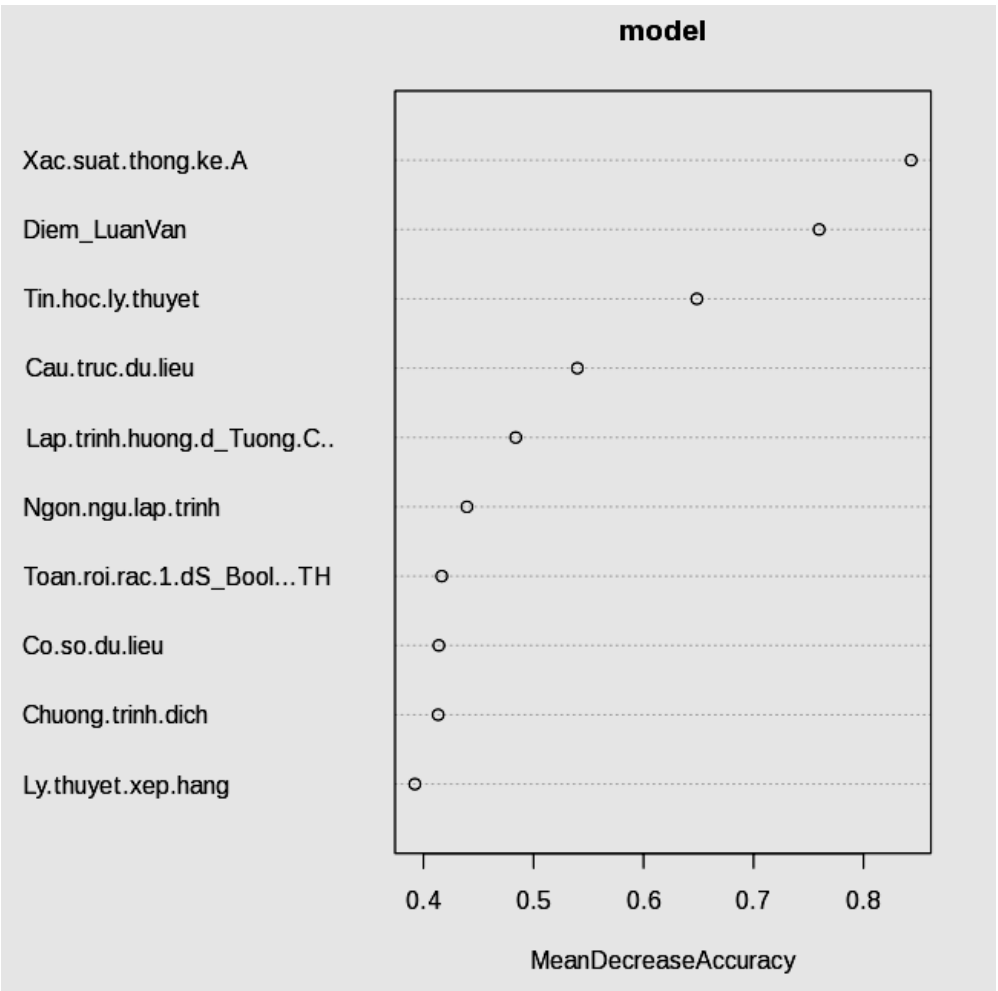
Hình 6: Kết quả học tập thu được sau bước tiền xử lý

Tiếp theo chúng tôi bỏ cột chứa các môn học mang tính chất điều kiện trong chương trình đào tạo như: giáo dục thể chất, giáo dục quốc phòng. Các môn học này sinh viên chỉ cần học đạt (hoàn thành), bởi vì kết quả của các môn học này không được sử dụng để đánh giá phân loại kết quả học tập. Trường (cột) cuối cùng là điểm trung bình tốt nghiệp được làm tròn đến phần nguyên được dùng làm cột loại (lớp), cột này có các giá trị từ 5, 6, 7, 8, 9 đến 10.

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 kết quả học tập của sinh viên.

4.2 Xây dựng mô hình rừng ngẫu nhiên và rút trích môn học quan trọng

Chương trình xử lý của chúng tôi dựa trên gói chương trình rừng ngẫu nhiên randomForest cung cấp sẵn trong môi trường ngôn ngữ R (Ihaka & Gentleman, 1996). Tiến hành xây dựng mô hình thực nghiệm rừng ngẫu nhiên với 200 cây quyết định, thực hiện lấy ngẫu nhiên 50 thuộc tính để tính phân hoạch tại mỗi nút. Quá trình rút trích đặc trưng (môn học) như mô tả ở phần trước. Sau đó sắp xếp thứ tự quan trọng các môn học giảm dần. Kết quả rút trích lấy 10 môn học có tính quan trọng nhất (ảnh hưởng đến kết quả đầu ra của sinh viên CNTT), được trình bày trong Hình 7.



Hình 7: Top 10 môn học quan trọng

Trong top 10 môn học quan trọng nhất được rút trích từ mô hình học rừng ngẫu nhiên, chúng ta có thể thấy rằng các môn học phân bổ vào 3 nhóm như sau.

– Nhóm các môn học đại cương: xác suất thống kê, toán rời rạc, vật lý lượng tử;

– Nhóm môn học cơ sở ngành: Tin học lý thuyết, Cơ sở dữ liệu, Ngôn ngữ lập trình, Cấu trúc dữ liệu, Lập trình hướng đối tượng;

– Nhóm môn học chuyên ngành: luận văn tốt nghiệp, chương trình dịch.

Theo như kết quả trình bày thì đây là những môn học đáng quan tâm nhiều nhất, vì kiến thức các môn học này ảnh hưởng đến chất lượng đào tạo của sinh viên ngành CNTT.

Để kiểm chứng kết quả rút trích đặc trưng (môn học) quan trọng, chúng tôi đánh giá lại hiệu quả của phân lớp dữ liệu, lần lượt trên tập dữ liệu với tập đầy đủ 249 đặc trưng (môn học) và tập dữ liệu chỉ với top 10 đặc trưng (môn học) quan trọng vừa rút trích. Kết quả ước tính tỉ lệ lỗi trên tập dữ liệu với đầy đủ 249 đặc trưng là 21,14%, trong khi tỉ lệ lỗi trên tập dữ liệu với top 10 đặc trưng chỉ là 20,82%. Điều này cho thấy rằng ta chỉ cần xác định được 10 môn học quan trọng này là có thể phân loại được kết quả học tập của sinh viên.

Tóm lại, với kết quả trên, nghiên cứu đã phát hiện ra những môn học quan trọng ảnh hưởng đến kết quả xếp loại tốt nghiệp của sinh viên ngành CNTT. Hay nói cách khác, sinh viên học tốt các môn học này thì kết quả xếp tốt nghiệp sẽ tốt. Những môn học này có thể sử dụng để phân loại kết quả học tập của sinh viên tốt nghiệp.

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày một tiếp cận khai mỏ dữ liệu để phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của sinh viên ngành CNTT tại Trường ĐHCT. Các bước thực hiện bao gồm sưu tập dữ liệu học tập của sinh viên tốt nghiệp ngành CNTT, sau đó thực hiện bước tiền xử lý dữ liệu để có thể huấn luyện mô hình rừng ngẫu nhiên cho phép rút trích các môn học quan trọng. Kết quả thu được sau khi rút trích các môn học quan trọng, có thể cung cấp thông tin hữu ích cho các nhà quản lý giáo dục, giảng viên, sinh viên trong việc tổ chức giảng dạy để nâng cao hiệu quả đào tạo.

Trong tương lai, chúng tôi dự định mở rộng nghiên cứu và phát triển cho các ngành đào tạo khác như kinh tế hay môi trường. Ngoài ra, cần phải tham khảo thêm ý kiến của các chuyên gia quản lý giáo dục, người sử dụng lao động để góp phần nâng cao độ tin cậy trong việc rút trích học phần quan trọng.

TÀI LIỆU THAM KHẢO

1. R. Agrawal, T. Imielinski and A. Swami.: Mining Associations between Sets of Items in Massive Databases. in proc. of ACM-SIGMOD International Conference on Management of Data, Washington, USA, pp. 207-216 (1993).
2. L. Breiman, J. Friedman, R. Olshen, C. Stone.: *Classification and Regression Trees*. Chapman & Hall (1984).
3. Breiman, L.: Bagging predictors. *Machine Learning* 24(2):123–140 (1996).
4. Breiman, L.: Random forests. *Machine Learning* 45(1): 5–32 (2001).
5. R. Bukralia, A-V. Deokar, S. Sarnikar, M. Hawkes.: Using Machine Learning Techniques in Student Dropout Prediction. Chapter 7 in *Cases on Institutional Research Systems*, Hansel Burley Eds., IGI Global, pp. 117-131 (2012).
6. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth.: From Data Mining to Knowledge Discovery in Databases. in *AI Magazine*, 17(3): 37-54 (1996).
7. T. Hastie, J-H. Friedman, R. Tibshirani.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer (2001).
8. R. Ihaka and R. Gentleman. A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3): 299-314 (1996).
9. J. MacQueen.: Some methods for classification and analysis of multivariate observations. in proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, *University of California Press* Vol.1, pp. 281-297 (1967).
10. T-N. Nguyen, L. Drumond, T. Horváth, L. Schmidt-Thieme.: Multi-Relational Factorization Models for Predicting Student Performance. in proc. of the KDD 2011 Workshop on Knowledge Discovery in Educational Data (2011).
11. A-K. Pal, S. Pal.: Analysis and Mining of Educational Data for Predicting the Performance of Students. *in International Journal of Electronics Communication and Computer Engineering* Vol.4(5): 2278-4209 (2013).
12. J. Pearl.: Bayesian Networks: a Model of Self-Activated Memory for Evidential Reasoning. in proc. of Cognitive Science Society, UC Irvine, pp. 329-334 (1985).

13. J.R. Quinlan.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA (1993).
14. V. Vapnik.: *The Nature of Statistical Learning Theory*. Springer-Verlag (1995).
15. L.A. Zadeh: Fuzzy sets. *Information and Control*, 8(3): 338–353 (1965).
16. Lê Thanh Minh.: Khai khoáng điểm thi tốt nghiệp phục vụ đánh giá phân loại học sinh. Luận văn Thạc sĩ. Đại học Khoa học Tự nhiên TP.HCM (2002).
17. Nguyễn Quốc Thông.: Phát triển một số ứng dụng khai thác dữ liệu vào giáo dục đào tạo. Luận văn Thạc sĩ. Đại học Khoa học Tự nhiên TP.HCM (2002).
18. Nguyễn Thái Nghe.: Một phân tích giữa các kỹ thuật trong dự đoán kết quả học tập. Kỷ yếu Hội thảo quốc gia lần thứ 10 về công nghệ thông tin, trang 19-31 (2007).
19. Phan Đình Thế Huân.: Nghiên cứu và ứng dụng phương pháp khai mỏ luật kết hợp trên dữ liệu giáo dục. Luận văn Thạc sĩ. Đại học Khoa học Tự nhiên TP.HCM (2009).
20. Nguyễn Thị Vân Hảo.: Xây dựng hệ thống dự đoán kết quả tốt nghiệp phổ thông trung học. Luận văn Thạc sĩ. Đại học Lạc Hồng, Đồng Nai (2011).
21. Nguyễn Đăng Nhượng: Khai phá dữ liệu về kết quả học tập của học sinh trường Cao đẳng nghề Văn Lang Hà Nội. Luận văn Thạc sĩ. Đại học Công nghệ, ĐHQGHN (2012).