APPLYING DECISION TREE ALGORITHM TO RECOMMEND COLLEGE MAJOR FOR HIGH SCHOOL STUDENTS

ỨNG DỤNG THUẬT TOÁN CÂY QUYẾT ĐỊNH ĐỂ TƯ VẤN CHỌN NGÀNH CHO HỌC SINH PHỔ THÔNG

Từ Thị Bích Hồng[1], Phạm Xuân Hậu[2]
[1] Trường THPT Phan Đình Phùng
[2] Trường Đại học Quảng Bình

ABSTRACT: *The choosing a college major is very important problem for high school students. The right selecting college major could help students for the career orientation in order to be suitable with their passion, opportunities and potential positions. Thus, they need recommendation and orientation from other to choose suitable major when submitting documents to colleges. In this paper, we apply the decision tree algorithm based on high school transcript of students to generate the set of rules and input them to recommendation system in order to suggest suitable college major.*

Keywords*: Subject, transcript, decision tree, attributes, college major.*

TÓM TẮT: *Việc chọn ngành đối với học sinh phổ thông là bài toán rất quan trọng. Việc chọn đúng ngành học giúp các em định hướng nghề nghiệp phù hợp với khả năng, đam mê và cơ hội của bản thân. Do vậy, các em cần có sự tư vấn, định hướng để có thể lựa chọn cho mình ngành nghề phù hợp khi đăng ký xét tuyển vào các trường đại học, cao đẳng. Trong bài báo này, chúng tôi áp dụng kỹ thuật cây quyết định dựa vào dữ liệu điểm các môn học của học sinh để đưa ra các tập luật suy luận, từ đó đưa vào hệ thống để tư vấn cho các em ngành học phù hợp khi chọn các ngành học ở các trường đại học, cao đẳng.*

Từ khóa*: Môn học, điểm môn học, cây quyết định, thuộc tính, ngành học.*

1. ĐẶT VẤN ĐỀ

Nghề nghiệp chính là phương tiện để đảm bảo vật chất cũng như tinh thần của con người, mỗi cá nhân đều phải lựa chọn cho mình một ngành nghề nhất định để tồn tại và phát triển. Đây không chỉ là phương thức sinh tồn mà còn là nơi mỗi người thực hiện mơ ước, lý tưởng của mình đồng thời góp phần vào sự phát triển của quê hương đất nước. Sự phát triển không ngừng của xã hội tạo ra sự phong phú về nghề nghiệp, tạo cho người lao động nói chung và học sinh Trung học phổ thông (THPT) nói riêng có nhiều cơ hội để tìm kiếm việc làm.

Việc chọn đúng ngành học giúp các em định hướng đi phù hợp với khả năng của bản thân mình và tránh khỏi những lựa chọn vội vàng để rồi phải bỏ lỡ rất nhiều cơ hội mà đáng ra nếu chọn và định hướng sớm và đúng các em sẽ thành công. Bên cạnh đó, việc chọn đúng ngành học trong bối cảnh kinh tế, xã hội hiện nay nhằm tránh việc có nhiều cá nhân lựa chọn sai nghề sẽ dẫn tới giảm sút chất lượng đào tạo, gây lãng phí cho công tác đào tạo và đào tạo lại. Chất lượng nguồn nhân lực sau đào tạo không đảm bảo dẫn tới năng suất lao động không cao, nảy sinh nhiều xáo trộn cho hoạt động

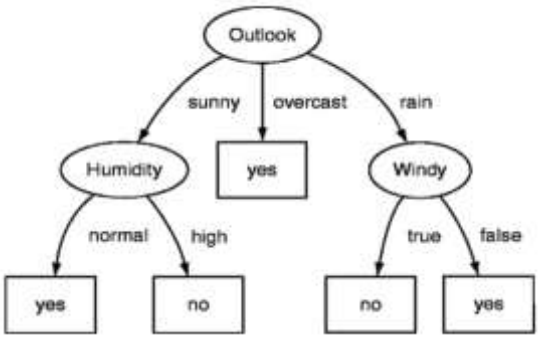
của các tổ chức, doanh nghiệp bởi các hiện tượng như: bỏ nghề, chuyển nghề,… Các doanh nghiệp mất thêm chi phí đào tạo và đào tạo lại cho đội ngũ của mình. Tuy nhiên, việc tiếp cận các thông tin về chọn ngành học cũng hạn chế, thiếu thông tin để lựa chọn và định hướng cho mình bởi do nhiều yếu tố khách quan, chủ quan khác nhau…

Chính vì vậy, công tác tư vấn ngành cho học sinh THPT là điều hết sức cần thiết. Trong bài báo này, chúng tôi thực hiện việc áp dụng thuật toán cây quyết định để phân lớp, xây dựng cây quyết định dựa vào hồ sơ điểm của các em học sinh và tạo ra các tập luật để làm cơ sở cho quá trình tư vấn. Dựa vào kết quả tư vấn đó sẽ giúp học sinh lựa chọn cho mình một ngành học phù hợp với điều kiện và năng lực của bản thân.

2. KỸ THUẬT CÂY QUYẾT ĐỊNH

2.1. Cây quyết định

Một cây quyết định là một mô hình logic được biểu diễn như một cây, cho biết giá trị của một biến mục tiêu có thể được dự đoán bằng cách dùng các giá trị của một tập các biến dự đoán [1,2,3,4]. Một số ưu điểm của cây quyết định như: xây dựng tương đối nhanh; đơn giản, dễ hiểu. Việc phân lớp dựa trên cây quyết định đạt được sự tương tự và đôi khi là chính xác hơn so với các phương pháp phân lớp khác. Hình 1 thể hiện một ví dụ về cây quyết định.



Hình 1. Ví dụ về cây quyết định

Trong đó:

▪ Gốc: Nút trên cùng của cây.

▪ Nút trong: Biểu diễn một kiểm tra trên một thuộc tính đơn.

▪ Nhánh: Biểu diễn các kết quả của kiểm tra trên nút trong.

▪ Nút lá: Biểu diễn lớp hay sự phân lớp.

Để phân lớp mẫu dữ liệu chưa biết, giá trị các thuộc tính của mẫu được đưa vào kiểm tra trên cây quyết định. Mỗi mẫu tương ứng có một đường đi từ gốc đến lá và lá biểu diễn dự đoán giá trị phân lớp mẫu đó. Cây quyết định được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây.

2.2. Xây dựng cây quyết định

Quá trình xây dựng cây quyết định gồm hai giai đoạn:

- Giai đoạn phát triển cây quyết định: Giai đoạn này bắt đầu từ gốc, đến từng nhánh và phát triển quy nạp theo cách thức chia để trị cho tới khi đạt được cây quyết định với tất cả các lá được gán nhãn lớp.

- Giai đoạn cắt, tỉa bớt các cành nhánh trên cây quyết định: Giai đoạn này nhằm mục đích đơn giản hóa và khái quát hóa từ đó làm tăng độ chính xác của cây quyết định bằng cách loại bỏ sự phụ thuộc vào mức độ lỗi của dữ liệu đào tạo mang tính chất thống kê, hay những sự biến đổi mà có thể là đặc tính riêng biệt của dữ liệu đào tạo.

Quá trình học cây quyết định gồm có 3 giai đoạn:

+ Tạo cây: Sử dụng các thuật toán phân lớp để phân chia tập dữ liệu huấn luyện một cách đệ quy cho đến khi mọi nút lá đều thuần khiết, tức là nút mà tại đó tập mẫu tương ứng có cùng một giá trị trên thuộc tính quyết định. Sự lựa chọn các thuộc tính

trong quá trình xây dựng cây được dựa trên việc đánh giá lượng lợi ích thông tin tại mỗi thuộc tính đang xét.

+ Cắt tỉa cây: Sau khi tạo cây, cắt tỉa cây quyết định là việc làm rất cần thiết để khắc phục những khiếm khuyết của cây. Cắt tỉa cây là cố gắng loại bỏ những nhánh không phù hợp hay những nhánh gây ra lỗi.

+ Kiểm định cây kết quả: Để kiểm tra độ chính xác của cây trước khi đưa vào ứng dụng trong thực tế, ta cần phải đánh giá độ chính xác của cây từ đó đưa ra tiêu chí đánh giá độ tin cậy theo tỷ lệ phần trăm được dự đoán chính xác.

1. BÀI TOÁN CHỌN NGÀNH CHO HỌC SINH VÀ MỘT SỐ KHẢO SÁT VẤN ĐỀ NÀY TRÊN ĐỊA BÀN TỈNH QUẢNG BÌNH

2.1. Vấn đề chọn ngành của học sinh THPT trên địa bàn tỉnh Quảng Bình

Kết thúc đợt 1 đăng ký thi tốt nghiệp THPT và xét tuyển đại học năm 2020, toàn tỉnh có 55,31% trong tổng số học sinh đăng ký thi chỉ để xét tốt nghiệp. Sự mất cân đối trong cơ cấu chọn ngành hiện nay. Học sinh đua nhau thi vào những ngành liên quan đến tài chính, tiền tệ, và bỏ qua các ngành nghề thuộc về khoa học cơ bản hay kỹ thuật là tình trạng phổ biến diễn ra trong nhiều năm. Nếu trong tương lai, tình trạng này vẫn tiếp tục, sẽ gây mất cân đối ngành nghề nghiêm trọng. Có thể thấy, thay vì đổ xô vào đại học, nhiều học sinh lựa chọn học nghề để tăng cơ hội tìm kiếm việc làm.

Việc thống kê các ngành học, nhóm ngành học và mã định danh được quy định tại Thông tư số 24/2017/TT-BGDĐT ngày 10 tháng 10 năm 2017 của Bộ trưởng Bộ Giáo dục và Đào tạo về cấp danh mục giáo dục, đào tạo cấp IV trình độ đại học [5]. Kết quả dữ liệu thu thập và xử lý cho thấy tỷ lệ ở trên có thể thấy rằng việc lựa chọn hiện nay của học sinh THPT đang có sự tập trung vào các ngành đó Quản trị kinh doanh, Ngành Marketing, Nhóm ngành Công nghệ thông tin, Ngành Kế toán, (Các ngành này xếp ở các vị trí thứ 1, 2, 3, 4 của bảng xếp hạng mà học sinh THPT lựa chọn). Học sinh lựa chọn chủ yếu vẫn là bên khối ngành kinh doanh, kinh tế, và theo đánh giá của các em thì thi vào những ngành này sẽ dễ xin việc và đem lại thu nhập cao. Có thể nói các khối ngành kinh tế đã có độ nóng từ rất lâu, cho đến tận thời điểm này khi mà nền kinh tế gặp suy thoái, kinh tế đất nước phát triển khó khăn thì vẫn không ảnh hưởng nhiều đến sự lựa chọn ngành nghề của các em.

2.2. Bài toán chọn ngành

Bài toán chọn ngành học tại các trường đại học, cao đẳng được phát biểu như sau: dựa vào các thông tin đầu vào của học sinh THPT như khu vực, đối tượng (ưu tiên), điểm trung bình THPT, điểm thi tốt nghiệp và tổ hợp xét tuyển mà các em muốn lựa chọn để từ đó tư vấn ngành học phù hợp với học sinh.

Đối với bài toán này chúng tôi áp dụng kỹ thuật cây quyết định để tìm ra các luật dựa trên hồ sơ dữ liệu của các học sinh đã chọn ngành (đã xét tuyển vào các trường đại học, cao đẳng) từ đó áp dụng các luật đã có để kết luận ngành học phù hợp cho các học sinh hiện tại.

Trong tiếp cận này, chúng ta đề cập đến hồ sơ học sinh. Hồ sơ là các mô tả rõ ràng các giá trị của các thuộc tính đối với một người dùng cá nhân bao gồm các thuộc tính như điểm trung bình THPT, điểm ưu tiên, điểm các môn xét tuyển và tổ hợp xét tuyển. Để giải quyết bài toán trên cần phân tích số

liệu từ hồ sơ học sinh. Từ đó, trích rút các tập luật để đề xuất các giải pháp giúp nhà trường nâng cao chất lượng đào tạo và tư vấn cho học sinh lựa chọn ngành nghề phù hợp.

1. ÁP DỤNG KỸ THUẬT CÂY QUYẾT ĐỊNH CHO BÀI TOÁN CHỌN NGÀNH CỦA HỌC SINH THPT

3.1. Thu thập dữ liệu

Dữ liệu được trích rút từ kết quả học tập (3 năm) tại trường THPT Hoàng Hoa Thám, trường THPT Phan Đình Phùng năm học 2017-2018; 2018-2019; 2019-2020. Danh sách đăng ký nguyện vọng vào các trường ĐH, CĐ trên cả nước của học sinh trường THPT Hoàng Hoa Thám và học sinh trường THPT Phan Đình Phùng năm học 2019-2020. Kết quả khảo sát thực tế các ngành đã được chọn của các em học sinh cũ của trường.

Từ các bảng dữ liệu trên chúng tôi đã tổ chức lại thành một bảng dữ liệu gồm các trường (thuộc tính) với gần 500 tập mẫu sau:

Mã trường: 50 mã trường (MATRUONG)

Tổ hợp xét tuyển: 13 tổ hợp (TOHOP) A00, A01, B00, C00, C14, C15, C19, C20, D01, D07, D15, D78, D96.

Mã ngành: 97 mã ngành (MANGANH) thuộc 13 tổ hợp theo danh sách các mã tổ hợp và môn xét tuyển theo tổ hợp được quy định tại Công văn số 310/KTKĐCLGD-TS của Bộ Giáo dục và Đào tạo về mã hóa các tổ hợp môn thi và xét tuyển Đại học, Cao đẳng chính quy [6].

- Ưu tiên (UUTIEN): 02 khu vực tuyển sinh và các ưu tiên khác theo quy định trong Quy chế tuyển sinh đại học, cao đẳng.

Tổng điểm môn xét tuyển (TONGXT): Từ được chia thành các mức như sau:

Tổng điểm xét tuyển	Gán nhãn
TONGXT >=26	26_30
23 <= TONGXT < 26	23_26
20 <= TONGXT < 23	20_23
17 <= TONGXT < 20	17_20
14 <= TONGXT < 17	14_17
11 <= TONGXT < 14	11_14
9 <= TONGXT < 11	9_11

Trung bình điểm học tập tại trường THPT (TBTHPT): Từ 1 – 10 được chia làm 4 khoảng: YEU, TB, KHA, GIOI

- Kết quả: DAU, HONG

3.2. Tìm các tập luật

Thuật toán cây quyết định được sử dụng để giải quyết cho bài toán này là thuật toán ID3. Thuật toán ID3 biểu diễn các khái niệm (concept) ở dạng các cây quyết định (decision tree). Biểu diễn này cho phép chúng ta xác định phân loại của một đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó [1,2,3,4]. Thuật toán ID3 tiếp cận kỹ thuật tìm kiếm tham lam đối với tập dữ liệu. Quá trình tiền xử lý dữ liệu và sinh tập luật được thực hiện bằng công cụ Weka.

- Xây dựng nút theo chiến lược Top-Down, bắt đầu từ nút gốc.

- Ở mỗi nút, thuộc tính kiểm tra là thuộc tính có khả năng phân loại tốt nhất.

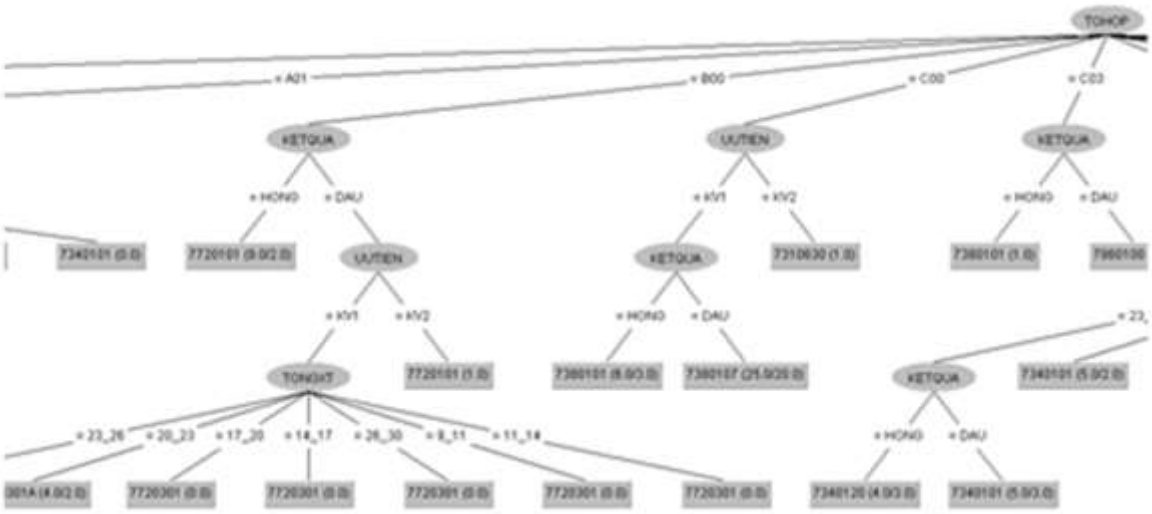
- Tạo mới một cây con của nút hiện tại cho mỗi giá trị có thể của thuộc tính kiểm tra, và tập dữ liệu đầu vào sẽ được tách ra thành các tập con tương ứng với các cây con vừa tạo.

- Mỗi thuộc tính chỉ được phép xuất hiện tối đa 1 lần đối với bất kỳ đường đi nào trong cây.

- Quá trình phát triển cây sẽ tiếp tục cho tới khi:

+ Cây quyết định phân loại hoàn toàn các dữ liệu đầu vào.

+ Tất cả các thuộc tính được sử dụng.

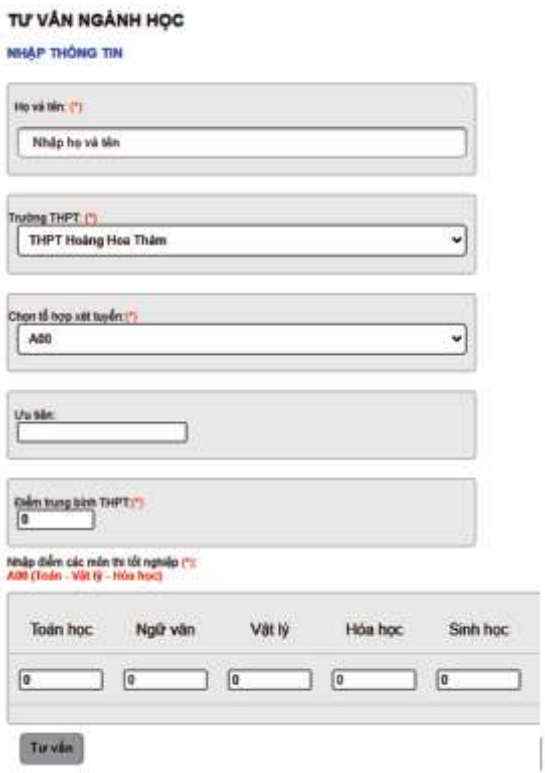


Hình 2. Cây quyết định

Một số tập luật được đề xuất sau khi chạy dữ liệu thu thập và áp dụng thuật toán ID3 như sau:

If TOHOP = A00 then
If TONGXT = 23_26 then
If (TBTHPT = GIOI) then
If (UUTIEN = KV1) then
MANGANH= 7810201
If (UUTIEN = KV2) then
MANGANH= 7480201
If TOHOP = B00 then
If UUTIEN = KV1 then
If (TONGXT = 23_26) then
If (TBTHPT = GIOI) then
MANGANH =7720301
If (TBTHPT = KHA) then
MANGANH = 7620110

Các tập luật sinh ra được đưa vào để cơ sở dữ liệu tập luật để thực hiện việc suy luận kết quả với đầu vào là dữ liệu của học sinh mới. Để được tư vấn học sinh nhập các thông tin theo yêu cầu của hệ thống như: Họ và tên, trường đang theo học, tổ hợp xét tuyển, điểm ưu tiên (nếu có), điểm trung



bình tổng kết THPT, các điểm các môn học tương ứng với các tổ hợp xét tuyển đã chọn. Đối với các mã tổ hợp xét tuyển có môn Khoa học xã hội (Lịch sử, Địa lý và Giáo dục công dân) và Khoa học tự nhiên (Vật lý, Hóa học và Sinh học) thì điểm sẽ được tính tổng 3 môn cộng lại và chia 3 để tính vào tổng điểm xét tuyển. Hình 3 mô tả giao diện của hệ thống các dữ liệu học sinh cần cung cấp cho hệ thống.

4. KẾT LUẬN

Sự phát triển không ngừng của xã hội tạo ra sự phong phú về nghề nghiệp, tạo điều kiện cho học sinh THPT có nhiều cơ hội để chọn ngành học phù hợp. Việc chọn đúng ngành học giúp các em định hướng đi phù hợp với khả năng của bản thân mình và tạo ra nguồn nhân lực tốt cho xã hội. Trong nghiên cứu này chúng tôi tập trung vào tìm hiểu và áp dụng thuật toán cây quyết định để phân lớp dữ liệu, xây dựng cây quyết định và sinh ra các tập luật. Từ kết quả các tập luật được sinh ra chúng tôi tiến hành xây dựng công cụ tư vấn cho học sinh THPT trong việc chọn ngành bằng các dữ liệu đầu vào được cung cấp và đã cho kết quả bước đầu chứng minh được tính khả thi của phương pháp tiếp cận.

TÀI LIỆU THAM KHẢO

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). *From Data Mining to Knowledge Discovery in Databases.* AAAI Press.

[2] Jiawei Han and Micheline Kamber (2011). *Data Mining: Concepts and Techniques*, 3rd Edition. Morgan Kaufmann Publishers.

[3] C. Heiner, N. Heffernan, T. Barnes. *Educational Data Mining (*2007). In Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education (AIED).

[4] Tom M. Mitchell (1997). *Machine Learning*, McGraw - Hill Science

[5] Thông tư số 24/2017/TT-BGDĐT, 10/102017 của Bộ trưởng Bộ Giáo dục và Đào tạo về cấp danh mục giáo dục, đào tạo cấp IV trình độ đại học.

[6] Công văn số 310/ KTKĐCLGD-TS của Bộ Giáo dục và Đào tạo. Bảng mã hóa các tổ hợp môn thi và xét tuyển Đại học, Cao đẳng chính quy.

Liên hệ:

TS. Phạm Xuân Hậu

Khoa Kỹ thuật - Công nghệ thông tin, Trường Đại học Quảng Bình

Địa chỉ: 312 Lý Thường Kiệt, thành phố Đồng Hới, tỉnh Quảng Bình

Email: pxhauqbu@gmail.com

Ngày nhận bài:

Ngày gửi phản biện:

Ngày duyệt đăng: