MÔ HÌNH PHÂN LOẠI SỬ DỤNG CÂY QUYẾT ĐỊNH ÁP DỤNG CHO HỆ THỐNG TUYỂN SINH CỦA TRƯỜNG ĐẠI HỌC

Đào Việt Anh
Khoa Công nghệ thông tin
Email: anhdv@dhhp.edu.vn

Ngày nhận bài: 09/11/2018
Ngày PB đánh giá: 27/01/2019
Ngày duyệt đăng: 08/02/2019

TÓM TẮT

Trong bài báo này, chúng tôi giới thiệu một kỹ thuật học máy có giám sát để xây dựng một cây quyết định cho hệ thống tuyển sinh của Trường đại học Hải Phòng. Mục tiêu chính là nhằm xây dựng được một mô hình phân loại hiệu quả với khả năng hạn chế lỗi cao và mức chính xác tương đối để cải thiện hiệu suất và hiệu quả của quá trình tuyển sinh. Điều này có nghĩa rằng công cụ lọc đã cải thiện hiệu suất và hiệu quả của quá trình tuyển sinh. Công cụ phân loại có chức năng lọc các ứng viên ở mức ban đầu để nhân viên tuyển sinh có thể tập trung vào các ứng viên triển vọng cao hơn nhằm đưa ra một lựa chọn tốt hơn. Vì vậy, khối lượng công việc của nhân viên hành chính được giảm bớt đi nhiều nên họ có thể thực hiện công việc lựa chọn tốt hơn.

Từ khóa: Khai phá dữ liệu, cây quyết định, đánh giá mô hình, học máy có giám sát, hệ thống tuyển sinh của trường đại học.

A DECISION TREE CLASSIFICATION MODEL FOR UNIVERSITY ADMISSION SYSTEM

ABSTRACT

This paper aims at introducing a supervised learning technique of building a decision tree for HaiPhong University admission system. The main object is to build an efficient classification model with high recall under moderate precision to improve the system. We used ID3 algorithm for decision tree construction. The final model is evaluated using the common evaluation methods. This means that the filtering tool has improved the efficiency and effectiveness of the admission process. The sorting tool has the ability to filter candidates at the initial level so that recruiters can focus on higher prospects in order to make a better choice. Therefore, the workload of administrative staff is reduced as they can conduct the selection better.

Keyword: Data mining, Decision tree, Model evaluation, Supervised learning, University Admission System.

I. ĐẶT VẤN ĐỀ

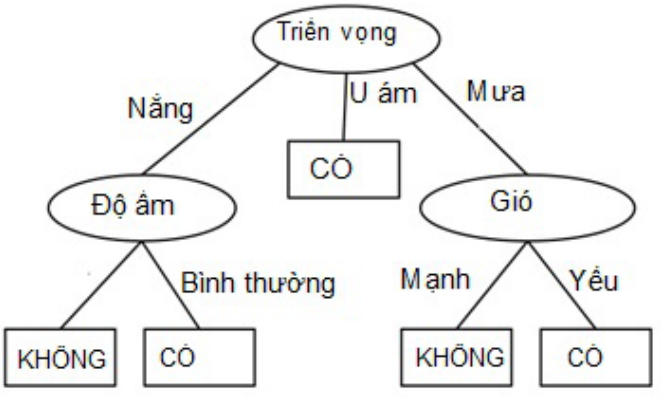
Khai phá dữ liệu nhằm tìm hiểu về những xu hướng chưa được biết đến, là một thành tố then chốt trong toàn bộ quá trình khám phá tri thức trong cơ sở dữ liệu. Trong kỷ nguyên máy tính ngày nay, những cơ sở dữ liệu này chứa những khối lượng thông tin khổng lồ. Khả năng tiếp cận và sự phong phú của khối thông tin này khiến vấn đề khai phá dữ liệu trở nên ngày càng quan trọng và cấp thiết [2].

Khai phá dữ liệu bao gồm nhiều phương pháp và kỹ thuật, nhưng chủ yếu ta có thể chia chúng thành hai loại: kiểm chứng và khai phá. Trong các phương pháp theo hướng kiểm chứng, hệ thống xác thực giả thiết đầu vào của người dùng như mức độ phù hợp, kiểm định giả thiết và kiểm định ANOVA. Mặt khác, các phương pháp theo hướng khai phá lại tự động tìm kiếm những quy tắc mới và xác định xu hướng trong dữ liệu. Các phương pháp theo hướng khai phá bao gồm kỹ thuật tạo cụm, phân loại và hồi quy.

Các phương pháp học máy có giám sát nhằm mục đích nhằm khai phá mối quan hệ giữa các thuộc tính đầu vào và thuộc tính đầu ra. Sau khi mô hình được xây dựng, ta có thể sử dụng mô hình đó để dự đoán giá trị của thuộc tính đầu ra đối với một dữ liệu đầu vào mới. Có hai nhóm mô hình có giám sát chính: mô hình phân loại (là mối quan tâm chính của chúng tôi trong bài viết này) và mô hình hồi quy. Mô hình phân loại xây dựng một bộ phân loại để ánh xạ không gian đầu vào (các đặc điểm) vào một trong các lớp định sẵn. Ví dụ, bộ phân loại có thể được sử dụng để phân loại các đối tượng trong một cảnh vật ngoài trời như người, phương tiện, cây hay tòa nhà. Trong khi đó, mô hình hồi quy ánh xạ không gian đầu vào với miền giá trị thực. Ví dụ, ta có thể dựng một mô hình hồi quy để dự đoán giá nhà dựa vào các đặc điểm như diện tích, số phòng, diện tích vườn…

Trong khai phá dữ liệu, cây quyết định (còn được gọi là Cây phân loại) là một mô hình dự đoán có thể được sử dụng để biểu diễn mô hình phân loại. Các cây phân loại có vai trò hữu dụng như một kỹ thuật khám phá và thường được sử dụng trong nhiều lĩnh vực như tài chính, marketing, y tế và kỹ thuật [1, 3, 7, 8]. Cây quyết định rất hay được được sử dụng trong khai thác dữ liệu nhờ tính đơn giản và dễ hiểu của chúng. Cây quyết định thường được biểu diễn về mặt đồ họa như một cấu trúc phân cấp, khiến chúng dễ diễn giải hơn các kỹ thuật khác. Cấu trúc này chủ yếu gồm có một nút bắt đầu (gọi là gốc) và nhóm các cành (nhánh hay điều kiện) dẫn đến các nút khác cho tới khi ta đến được nút lá chứa quyết định cuối cùng của tuyến này. Cây quyết định là một mô hình tự khám phá bởi cách biểu diễn cây rất đơn giản. Mỗi nút trong kiểm tra một thuộc tính, trong khi mỗi cành (nhánh) thì tương ứng với giá trị của thuộc tính (hay khoảng giá trị). Cuối cùng, mỗi lá được đặt cho một (cách) phân loại.

Hình 1 nêu ví dụ về một cây quyết định đơn giản cho phân loại "Chơi tennis". Cây đơn thuần quyết định xem có chơi tennis hay không (có các lớp Có hoặc Không) dựa vào ba thuộc tính thời tiết là triển vọng, gió và độ ẩm [5].



Hình 1: Ví dụ về cây quyết định.

Như minh họa trong Hình 1, nếu ta có một xu hướng mới với các thuộc tính triển vọng là "Mưa" và gió "Mạnh", vậy thì ta sẽ quyết định không chơi tennis bởi tuyến bắt đầu từ nút gốc sẽ kết thúc ở lá quyết định thuộc lớp "KHÔNG".

Trong bài viết này, chúng tôi giới thiệu một kỹ thuật học máy có giám sát để xây dựng mô hình cây quyết định cho hệ thống tuyển sinh của Trường đại học Hải Phòng nhằm cung cấp một công cụ lọc giúp cải thiện hiệu quả và hiệu suất của quá trình tuyển sinh. Hệ thống tuyển sinh gồm có một cơ sở dữ liệu chứa các hồ sơ về thông tin của học viên đăng ký và trạng thái của học viên là bị từ chối hay được chấp nhận tuyển vào học tại trường. Ta phải phân tích những hồ sơ này để xác định mối quan hệ giữa dữ liệu của người đăng ký với trạng thái thu tuyển cuối cùng.

Bài viết này được chia thành năm phần. Ở phần 2, chúng tôi trình bày mô hình cây quyết định. Phần 3 nêu sơ bộ về các phương pháp thường được sử dụng để đánh giá mô hình cây này. Ở phần 4, chúng tôi trình bày và phân tích kết quả thực nghiệm theo kết quả của cây quyết định và quan điểm của hệ thống tuyển sinh này. Cuối cùng, phần kết luận cho nghiên cứu này được trình bày trong Phần 5.

II. MÔ HÌNH CÂY QUYẾT ĐỊNH

Cây quyết định là một công cụ phân loại được biểu diễn dưới dạng một phân hoạch của không gian đầu vào dựa trên các giá trị thuộc tính. Như đã trình bày ở trước, mỗi nút trong của cây sẽ tách không gian trường hợp thành hai hoặc nhiều không gian con theo hàm nhất định của giá trị thuộc tính đầu vào. Mỗi lá được gán với một lớp biểu diễn giá trị mục tiêu thích hợp hoặc giá trị xảy ra thường xuyên nhất.

Các trường hợp được phân loại bằng cách đi xuyên qua cây từ nút rễ xuống lá theo kết quả của các nút kiểm định trên đường đi này. Khi đó, mỗi đường đi có thể được biến thành một quy tắc bằng cách ghép các kiểm định dọc theo đường đi này. Ví dụ, một trong các đường đi ở Hình 1 có thể được biến thành quy tắc sau: "Nếu Triển vọng trời Nắng hoặc Độ ẩm là Bình thường thì chúng ta có thể chơi tennis".

Có nhiều thuật toán được đề xuất để cây quyết định học hỏi từ một tập dữ liệu cho trước, song chúng tôi sẽ sử dụng thuật toán ID3 nhờ tính đơn giản và dễ triển khai của thuật toán này. Trong phần này, chúng

tôi sẽ bàn về thuật toán ID3 trong xây dựng cây quyết định và một số hàm thường được sử dụng để tách không gian đầu vào.

A. Thuật toán ID3

ID3 là một thuật toán học máy sử dụng cây quyết định do Quinlan [6] phát triển. Đầu vào là 1 tập dữ liệu huấn luyện bao gồm các mẫu dữ liệu. Mỗi mẫu dữ liệu bao gồm 1 tập các giá trị ứng với các thuộc tính. Ví dụ: bảng mẫu dữ liệu dưới thể hiện đội bóng có chơi hay không tương ứng với các kiểu thời tiết.

Bảng 1: Ví dụ mẫu dữ liệu

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Thuật toán này đơn giản sử dụng kiểu tìm kiếm từ trên xuống đối với tập các thuộc tính đầu vào cần được kiểm định tại mọi nút trên cây. Thuộc tích có độ phân chia tốt nhất theo hàm tiêu chí phân chia được sử dụng để tạo nút hiện tại. Quá trình này được lặp lại tại mọi nút cho tới khi một trong các điều kiện sau được đáp ứng:

Bao gồm mọi thuộc tính dọc theo đường dẫn này.

Các ví dụ rèn luyện hiện tại ở nút này có cùng giá trị mục tiêu.

Hình 2 thể hiện mã giả cho thuật toán ID3 khi xây dựng cây quyết định cho một tập rèn luyện (*S*), tập đặc điểm đầu vào (*F*), đặc điểm đầu ra (*c*) và một tiêu chí phân chia (*SC*) nào đó.

B. Tiêu chí phân chia

Thuộc tính ID3 sử dụng một hàm tiêu chí phân chia nào đó nhằm chọn thuộc tính tốt nhất để tách. Để xác định tiêu chí này, trước tiên ta cần xác định chỉ số entropy đo lường mức độ *pha tạp* của một tập dữ liệu được gắn nhãn nhất định.

Đối với một tập dữ liệu được gắn nhãn S cho trước với một số ví dụ có n (giá trị mục tiêu) lớp $\{c_1, c_2, ..., c_n\}$, ta có thể định nghĩa chỉ số entropy (E) như trong (1).

$$E(S)=\sum_{i=1}^{n} p_1 * \log(p_i), p_i = \frac{|S_{C_i}|}{|S|} \quad (1)$$

Trong đó S_{c_i} là tập con gồm các ví dụ có giá trị mục tiêu bằng c_i. Entropy (E) có giá trị tối đa nếu tất cả các lớp có cùng xác suất (xảy ra).

*ID*3(S, F, c, SC)

Đầu ra: Cây quyết định *T*

Tạo một cây quyết định *T* với một nút *gốc* duy nhất

IF không có thêm phân chia (S) **THEN**

Đánh dấu *T* là *lá* với giá trị phổ biến nhất của *c* lấy làm nhãn.

ELSE

$\forall f_i \in F$ tìm f có $SC(f_i, S)$ tốt nhất

Gắn nhãn t là f

FOR mỗi giá trị v_j bằng f

Đặt $Subtree_j = ID3\left(S_{f=v_j}, F-\{f\}, c, SC\right)$

Nối nút *t* với $Subtree_j$ với nhãn cạnh là dv_j

Hình 2. Thuật toán ID3

1) Độ tăng thông tin(thu thập được)

Để chọn thuộc tính tốt nhất nhằm tách một nút nhất định, ta có thể sử dụng thước đo độ tăng thông tin giả sử là Gain (S, A) của một thuộc tính A, bằng một tập ví dụ S. Độ tăng thông tin được định nghĩa trong (2).

$$Gain(S,A) = E(S) - \sum_{v \in V(A)} \frac{|S_{A=v}|}{|S|} E(S_{A=V}) \quad (2)$$

Trong đó E(S) là chỉ số entropy của tập dữ liệu S, V(A) là tập tất cả các giá trị của thuộc tính A.

2) Hệ số tăng

Một thước đo khác có thể được sử dụng như một tiêu chí phân chia đó là hệ số tăng. Đó đơn giản là hệ số giữa giá trị độ tăng thông tin Gain(S, A) và một giá trị khác, thông tin phân chia, SInfo(S, A), được định nghĩa trong (3).

$$SInfo(S,A) = \sum_{v \in V(A)} \frac{|S_{A=V}|}{|S|} * \log \frac{|S_{A=V}|}{|S|} \quad (3)$$

3) Thuật toán Relief

Kira và Rendell đã đưa ra đề xuất về thuật toán Relief ban đầu nhằm ước tính chất lượng của các thuộc tích theo việc giá trị của chúng khác biệt tốt như thế nào giữa các ví dụ gần giống nhau [4]. Các bước của thuật toán được nêu trong Hình 3, trong đó hàm diff tính toán sự khác nhau giữa cùng một giá trị thuộc tính (A) trong hai trường hợp khác nhau là I1 và I2 như trong (4).

$$diff(A, I_1, I_2) = \begin{cases} 0 & I_1[A] = I_2[A] \\ 1 & \text{trong những trường hợp khác nhau} \end{cases} \quad (4)$$

Relief

Đầu vào: Tập rèn luyện *S* có *N* ví dụ và *K* thuộc tính

Đầu ra: Véc-tơ trọng số *W* cho tất cả thuộc tính *A*

Đặt tất cả trọng số *W* [1..*K*] = 0

FOR i = 1 TO N

Chọn ví dụ ngẫu nhiên **R.**

Tìm lần *trúng* gần nhất **H** (trường hợp cùng lớp).

Tìm lần *trượt* gần nhất **M** (trường hợp khác lớp).

FOR A = 1 TO K

$$W[A] = W[A] - \frac{diff(A,R,H)}{N} + \frac{diff(A,R,M)}{N}$$

END; RETURN *W.*

Hình 3. Thuật toán Relief

III. ĐÁNH GIÁ MÔ HÌNH

Xét một bài toán lớp nhị phân (tức là chỉ có hai lớp: positive- dương tính, lớp còn lại là negative – âm tính), dữ liệu đầu ra của một mô hình phân loại là số trường hợp đúng và sai so với lớp đã biết trước đó của chúng. Những số này được lập thành đồ thị trong ma trận lỗi như thể hiện trong Bảng 2. Cách đánh giá này thường được áp dụng cho các bài toán phân lớp có hai lớp dữ liệu. Cụ thể hơn, trong hai lớp dữ liệu này có một lớp nghiêm trọng hơn lớp kia và cần được dự đoán chính xác. Ví dụ, trong bài toán xác định có bệnh ung thư hay không thì việc không bị sót quan trọng hơn là việc chẩn đoán nhầm âm tính thành dương tính.

Bảng 2. Ma trận lỗi (Bài toán lớp nhị phân)

Lớp thực	**Lớp dự đoán**		
	Dương tính	**Âm tính**	
Dương tính	TP	FN	CN
Âm tính	FP	TN	CP
	RN	**RP**	**N**

Như thể hiện trong bảng 1, TP (*True Positive*) là số trường hợp được dự đoán đúng là lớp *dương tính*. FP (*False Positive*) biểu diễn các trường hợp được dự đoán là *dương tính* trong khi thực sự thì lại thuộc lớp *âm tính*. Điều này cũng áp dụng với TN (*True Negative*) và FN *(False Negative*). Các tổng hàng CN và CP thể hiện số trường hợp *thực sự âm tính và thực sự dương tính;* các tổng cột RN và RP là số trường hợp *được dự đoán là âm tính và dương tính*. Cuối cùng, N là tổng số trường hợp trong tập dữ liệu.

Có nhiều biện pháp đánh giá được sử dụng để đánh giá hiệu quả của một công cụ phân loại căn cứ vào ma trận lỗi của công cụ ấy sau khi kiểm định. Chúng tôi sẽ thảo luận chi tiết hơn về một số biện pháp thường được sử dụng ở phần sau trong thử nghiệm của mình.

Độ chính xác của phân loại (***Acc***) là thước đo hay được sử dụng nhất để đánh giá tính hiệu quả của một công cụ phân loại theo *tỷ lệ phần trăm các trường hợp dự đoán đúng như trong (5).*

$$Acc = \frac{TP + TN}{N} \quad (5)$$

Mức ghi nhớ (***R- Recall***) là tỷ lệ phần trăm các trường hợp thuộc lớp dương tính và được dự báo là duong tính và Mức chính

xác (***P***) là tỷ lệ phần trăm các các trường hợp thuộc lớp dương tính được dự báo đúng. Các thước đo này căn cứ vào dữ liệu của ma trận lỗi:

$$R = \frac{TP}{CN} \text{ và } P = \frac{TP}{RN} \quad (6)$$

Cả Precision và Recall đều là các số nhỏ hơn hoặc bằng một. Precision cao đồng nghĩa với việc độ chính xác của các điểm tìm được là cao. Recall cao đồng nghĩa với tỉ lệ bỏ sót các điểm thực sự *dương tính* là thấp.

Mức chính xác và mức ghi nhớ có thể được kết hợp lại với nhau để hợp thành một thước đo khác gọi là "***F-measure***" như thể hiện trong (7). Một hằng số β được sử dụng để kiểm soát sự đánh đổi giữa các giá trị ghi nhớ và mức chính xác. Giá trị thường được sử dụng nhất cho β là 1, biểu diễn thước đo F_1.

$$F_\beta = \frac{\left(1+\beta^2\right) * P * R}{\left(\beta^2 * P\right) + R} \quad (7)$$

Đối với tất cả các thước đo xác định ở trên, khoảng giá trị của chúng dao động từ 0 đến 1. *Đối với một công cụ phân loại tốt, giá trị của từng thước đo nên gần bằng 1.*

IV. THỬ NGHIỆM

A. Tập dữ liệu

Hệ thống tuyển sinh của Trường đại học Hải Phòng là một quá trình ra quyết định phức tạp, không chi đơn thuần là so khớp điểm kiểm tra với các yêu cầu tuyển sinh mà còn bởi nhiều lý do. Thứ nhất, trường đại học có nhiều chi nhánh như các trường liên kết ở Hải Dương hay Thái Bình áp dụng cho cả hai nhóm, thí sinh nam và nữ. Thứ hai, số người đăng ký mỗi năm là rất lớn, do đó cần một tiêu chí lựa chọn phức tạp phụ thuộc vào thứ hạng ở trung học và khu vực/thành phố của người đăng ký.

Trong bài viết này, chúng tôi được cấp một tập dữ liệu mẫu từ cơ sở dữ liệu của hệ thống của trường, trong đó biểu diễn thông tin của thí sinh đăng ký và trạng thái bị từ chối hoặc được chấp nhận thu tuyển vào học tại trường đại học của thí sinh trong ba năm liên tiếp (2015, 2016 và 2017). Tập dữ liệu gồm 80262 hồ sơ, trong khi mỗi hồ sơ biểu diễn một trường hợp với 4 thuộc tính và thuộc tính lớp có hai giá trị: Bị từ chối và Được chấp nhận. Các lớp được phân phối chiếm 53% tổng số hồ sơ đối với lớp "Bị từ chối" và 47% đối với lớp "Được chấp nhận". Bảng 2 thể hiện thông tin chi tiết về các thuộc tính của tập dữ liệu.

Tập dữ liệu được chia thành hai phần chính: tập dữ liệu huấn luyện chứa 51206 hồ sơ (khoảng 64%). và tập dữ liệu kiểm tra đánh giá mô hình chứa khoảng 29056 hồ sơ (khoảng 36%). Công cụ phân loại cây quyết định được cho học hỏi bằng cách sử dụng tập dữ liệu huấn luyện và hiệu quả của công cụ được đo lường trên các tập dữ liệu kiểm tra đánh giá chưa từng thấy trước đó.

Bảng 3: Tổng hợp các thuộc tính của tập dữ liệu

Thuộc tính	Giá trị có thể
Giới tính	Giới tính của sinh viên • Nam • Nữ
HSGrade	Điểm ở trung học • Giỏi: Điểm > 8.5 • Khá: 7.5<Điểm<8.5 • Trung bình: 6.5< điểm < 7.5 • Kém : điểm < 6.5
Vùng	Mã thành phố thuộc khu vực của thí sinh

B. Kết quả của mô hình cây quyết định

Mô hình cây quyết định được khởi tạo từ các hồ sơ trong tập dữ liệu rèn luyện bằng cách sử dụng công cụ khai thác dữ liệu Orange[9]. Các giá trị của ma trận lỗi được thể hiện trong bảng 4. Các giá trị của ma trận lỗi được khởi tạo bằng cách áp dụng cây quyết định lên các tập dữ liệu kiểm định

Bảng 4: Ma trận lỗi đã được kiểm định

Lớp thực	Lớp dự đoán		
	Được chấp nhận	**Bị từ chối**	
Được chấp nhận	12305	1538	**13843**
Bị từ chối	8484	6729	**15213**
	20789	**8267**	**29056**

Bảng 5. Các thước đo đánh giá mô hình

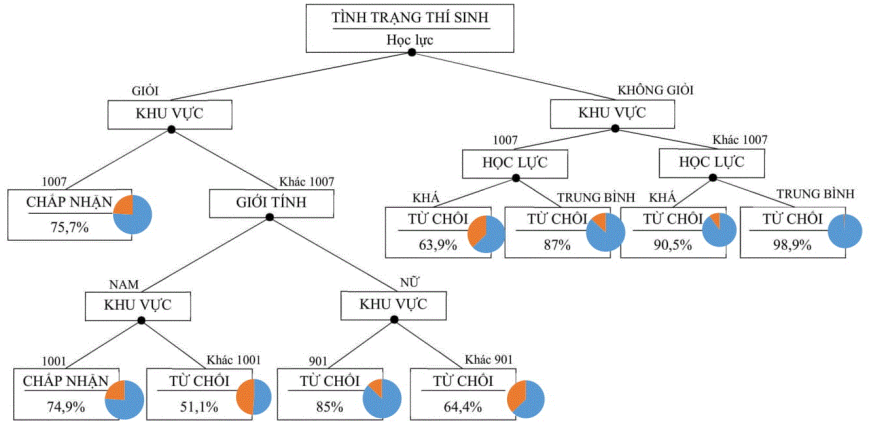
Giá trị đo	
Độ chính xác	$A_{cc} = \frac{12305 + 6729}{29056}$
Mức ghi nhớ	$R_{\text{Được chấp nhận}} = \frac{12305}{13843} = 0.889$ $R_{\text{Bị từ chối}} = \frac{6729}{5213} = 0.442$
Mức chính xác	$P_{\text{Được chấp nhận}} = \frac{12305}{20789} = 0.592$ $P_{\text{Bị từ chối}} = \frac{6729}{8267} = 0.834$
F_1 Độ đo	$F_{1\ \text{Được chấp nhận}} = \frac{2*0.592*0.889}{0.592+0.889} = 0.711$ $F_{1\ \text{Bị từ chối}} = \frac{2*0.834*0.442}{0.834+0.442} = 0.578$

Các thước đo đánh giá nêu trong bảng 5 cho thấy rằng công cụ phân loại đề xuất đã đạt được khả năng hạn chế lỗi cao, đổi lại là mức chính xác ở mức vừa phải. Điều này có nghĩa rằng công cụ lọc đã cải thiện hiệu suất và hiệu quả của quá trình tuyển sinh. Công cụ phân loại có chức năng lọc các thí sinh ở mức ban đầu để nhân viên tuyển sinh có thể tập trung vào các thí sinh triển vọng cao hơn nhằm đưa ra một lựa chọn tốt hơn. Vì vậy, khối lượng công việc của nhân viên hành chính được giảm bớt đi nhiều nên họ có thể thực hiện công việc lựa chọn tốt hơn. Trên thực tế, việc bỏ quên một số (tức là có mức ghi nhớ hơi thấp hơn 1) cũng không hẳn là điều tệ gì, bởi nhân viên hành chính không phải lúc nào cũng có thể xác định được ứng viên tốt nhất từ một nhóm nhiều thí sinh. Mặt khác, cũng các thước đo đó trong trường hợp lớp "Bị từ chối" đạt mức khoảng 0,58. Giá trị mức trung bình này

cho biết rằng hiệu quả của công cụ phân loại ở trên mức trung bình.

C. Cây quyết định và các quy tắc rút ra từ cây quyết định

Từ các yếu tố trên ta có thể đưa ra cây quyết định kết quả cuối cùng đối với 1 thí sinh như sau:



Hình 4. Cây quyết định kết quả thí sinh ứng tuyển

Một trong những ưu điểm chính của cây quyết định đó là cây có thể được giải thích như một tập quy tắc. Những quy tắc này được rút ra bằng cách đi xuyên qua cây, bắt đầu từ nút gốc cho tới khi đến một quyết định tại một lá. Những quy tắc này cũng giúp ta có một cái nhìn phân tích rõ ràng về hệ thống đáng xét. Trong trường hợp của chúng tôi, những quy tắc này sẽ giúp phòng hệ thống tuyển sinh hiểu được quy trình chung. Tập quy tắc suy ra được nêu trong bảng 6.

Bảng 6. Tập quy tắc từ cây quyết định

IF Khuvực= ”1007” **AND** HS Grade = ”Giỏi” **THEN** “Được chấp nhận” (75.7%)
IF Khuvực ≠ ”1007” **AND** HS_Grade = ”Giỏi” **AND** Giới tính = ” Nam” **AND** Khuvực = ”1001” **THEN** -‘Được chấp nhận” (74.9%)
IF Khuvực ≠ ” 1007” **AND** HS Grade = ”Giỏi” **AND** Giới tính = ”Nữ” **AND** Khuvực # ”901” **THEN** “Bị từ chối” (64.4%)
IF Khuvực ≠ ”1007” **AND** HS_Grade = ”Giỏi” **AND** Giới tính = ”Nữ” **AND** Khuvực= ”901” **THEN** “Bị từ chối” (85.0%)
IF Khuvực ≠ ”1007” **AND** HS Grade ≠ “Giỏi” **AND** HS Grade ≠ ”Khá” **THEN** “Bị từ chối” (98.9%)
IF Khuvực ≠ ”1007” **AND** HS_Grade = ”Giỏi” **AND** Giới tính = ”Nam” **AND** Khuvực ≠ ”1001 ” **THEN** “Bị từ chối” (51.1%)
IF Khuvực# ”1007” **AND** HS Grade ≠ ”Giỏi” **AND** HS Grade = ”Khá” **THEN** “Bị từ chối” (90.5%)
IF Khuvực= ” 1007” **AND** HS Grade ≠ *“Giỏi”* **AND** HS Grade ≠ ”Khá” **THEN** “Bị từ chối” (87.0%)
IF Khuvực= ” **1**007” **AND** HS_Grade ≠ “Giỏi” **AND** HS_Grade = ”Khá” **THEN** “Bị từ chối” (63.9%)

Như thể hiện trong bảng 6, mỗi quy tắc lại có tỷ lệ phần trăm số trường hợp được dự đoán bằng quy tắc này và theo lớp đó. Ta cũng có thể nhận thấy rằng chỉ có hai quy tắc dẫn đến trạng thái "Được chấp nhận". Trường hợp thứ nhất là khi mã vùng của thí sinh là "1007" (tức là khu vực thành phố "Hải Phòng") và điểm ở trung học của thí sinh là "Giỏi". Trường hợp thứ hai là khi sinh viên "Nam" từ vùng có mã "1001" (tức là khu vực lân cận thành phố "Hải Phòng") có điểm "Giỏi" ở trung học.

Sau khi sử dụng các thuật toán quyết định này thì lời khuyên dành cho bộ phân tuyển sinh trường Đại học Hải Phòng là nên tập trung vào các ứng viên có hộ khẩu gần Hải Phòng hay là các huyện vùng ven thành phố Hải Phòng thay vì các ứng viên ở các tỉnh xa. Đó là do các thí sinh này có xu hướng gắn bó với trường lâu hơn các thí sinh xa nhà do chi phí xa nhà cao và đặc tính địa phương của trường. Lưu ý này cũng hướng tới bộ phận tuyển sinh của trường là điều kiện tuyển sinh đầu tiên nên là Khu vực thay vì Điểm của thí sinh học ở bậc phổ thông.

V. KẾT LUẬN

Trong bài viết này, chúng tôi đã trình bày một mô hình phân loại hiệu quả bằng cách sử dụng cây quyết định cho phòng tuyển sinh của trường đại học. Kết quả thực nghiệm cho thấy rằng công cụ lọc đã cải thiện hiệu suất và hiệu quả của quá trình tuyển sinh. Quá trình phân loại này đạt được bằng cách sử dụng cây quyết định với khả năng hạn chế lỗi cao và mức chính xác tương đối. Chúng tôi đã thiết lập được các bộ quy tắc bằng cách sử dụng cấu trúc của cây quyết định và các bộ quy tắc này giúp cho việc lựa chọn thí sinh dễ dàng hơn.

TÀI LIỆU THAM KHẢO

1. J.Choand P.U.Kurup(2011), "Decision tree approach for classfication and dimensionality reduction of electronic nose data" , *Sensor & Actuators B Chemical, vol 160(1),542-548*
2. J.Han and M.Kamber,(2000*),"*Data mining: concepts and techniques", San Francisco, Morgan-Kaufrna.
3. H.S.OH and W.S.SEO,(2012*),"*Development of a Decision Tree Analysis model that predicts recovery from acute brain injury*", Japan Journal of Nursing Science, doi:10.1111/j 1742-7924-2012.00215.x.*
4. K. Kira and L.A. Rendeil, (1992),"A practical approach to feature selection", *In D. Sleeman and P.Edwards, edito, proceedings of international conference on Machine learning, pp 249-256, Morgan Kaufmann*
5. T. Michel, (1997), "Machine Learning", USA, Mc Graw Hill
6. J.R.Quinlan, (1986)," Introduction of Decision tree", *Machine Learning vol 1, pp 86-106.*
7. S.Sohn and J.Kim, (2012), "Decision tree – based technology credit scoring for start up firms, Korean case"*, Expert System with Applications vol 39(4), 4007-4012, doi 10.1016/j. eswa 2011.09.075*
8. G.Zhou and L.Wang,(2002),"Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation"*, Transportation research part C,21(1),287-305 doi: 10.1016/j trc.2011.10.007*
9. Orange Data mining tool: *http://orange .biolab.si.*