THIẾT KẾ VÀ ỨNG DỤNG CỦA CÂY QUYẾT ĐỊNH

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

Ngày nhận bài: 24/10/2017
Ngày PB đánh giá: 27/11/2017
Ngày duyệt đăng: 01/12/2017

TÓM TẮT

Cây quyết định là một trong những công cụ quan trọng nhất dùng để đưa ra những quyết định cho những tình huống không chắc chắn. Bài báo thảo luận quy trình thiết kế và xây dựng cây phân cấp quyết định với những thống kê cụ thể trong lĩnh vực có rất nhiều tính thiết thực trong đời sống như bảo hiểm và quy trình chấp thuận việc sản xuất phần mềm.

Từ khoá: Cây quyết định nhiều cấp, chính sách bảo hiểm, quy trình chấp thuận việc sản xuất phần mềm.

DESIGN AND APPLICATION OF DECISION TREE

ABSTRACT

Decision trees are one of the most important tools used to make decisions for uncertain situations. This paper discusses the process of designing and constructing decentralized trees with specific statistics in the field of real life such as insurance and the process of approving software production.

Keywords: Multilevel decision tree, Policy insurance, Software approval procedure.

---

1. GIỚI THIỆU

Để ra được quyết định trong những lĩnh vực phức tạp như chính sách bảo hiểm hay quy trình sản xuất phần mềm là một vấn đề khó khăn trong đó cây phân cấp quyết định có thể là một trong những giải pháp tối ưu.

Cây quyết định là một công cụ [1] sử dụng mô hình cấu trúc cây hoặc mô hình cây phân cấp quyết định [6]. Mục tiêu của cây quyết định là dự đoán bằng cách chia nhỏ cây theo các nhánh nhỏ hơn. Mỗi nhánh của cây quyết định thể hiện 1 khả năng có thể xảy ra của cây quyết định. Mục tiêu chính của cây quyết định là đưa ra được câu trả lời rõ ràng cho những trường hợp phức tạp, có quá nhiều lựa chọn hay không chắc chắn. Cây quyết định cho phép chúng ta mô hình hóa 1 tình huống phức tạp theo những giải pháp và định dạng nó một cách đơn giản và có thể hiểu được, đồng thời miêu tả được mối liên hệ giữa các quyết định khác nhau.

Có 3 loại nút trong cây quyết định:

- Nút gốc (nút quyết định)
- Nút trong (nút thay đổi có định hướng)

- Nút lá (nút kết quả).

Nút gốc đại diện cho vấn đề chính của một trường hợp không chắc chắn nào đó. Kết quả cuối cùng thường được trích ra từ những tính chất cơ bản của nút gốc. Nút này trong cây quyết định chúng ta sẽ quy ước là hình chữ nhật. Nút trong là các nút điều kiện, những nút này thường bao gồm các điều kiện đặc biệt và những nhánh từ các nút này cũng bao gồm các đầu ra tương ứng với điều kiện ấy. Nút trong chúng ta quy ước được vẽ bằng hình chữ nhật. Cuối cùng là nút lá, là nút chứa kết quả, bao gồm các quyết định về vấn đề. Những nút lá này quy ước được vẽ bằng hình tam giác.

2. NỘI DUNG NGHIÊN CỨU

2.1 Quá trình thiết kế cây quyết định

Thiết kế một cây quyết định T từ 1 bảng dữ liệu D bao gồm 4 quy trình tuân theo nguyên tắc chia để trị [7]. Một bảng dữ liệu được cho với cặp <A,R> trong đó A là tập hợp các thuộc tính còn R là tập hợp các bản ghi tương ứng với các thuộc tính đó.

Giả sử có 1 bảng dữ liệu:

$D=\{<A_1,R_1>,<A_2,R_2>,..,<A_i,R_i>\}$.

Quá trình thiết kế cây quyết định tuân theo các bước sau:

Bước 1. Nếu tất cả các bản ghi đều được gán với cùng 1 thuộc tính thì trả lại kết quả là nút lá, nút có cùng thuộc tính ấy.

Bước 2. Chọn vài điều kiện "t" bao gồm 2 hoặc nhiều hơn đầu ra, ví dụ như "t_1" đến "t_i" cho bản ghi thứ i.

Bước 3. Giờ tất cả bảng dữ liệu đã được chia thành tập hợp các bảng dữ liệu con $<D_1, D_2 ,...,D_i>$ trong đó D_i chứa đầu ra tương ứng với điều kiện t_i

Bước 4. Như chúng ta đã biết quy tắc chia để trị [7] , thực hiện đệ quy quá trình này từ tập con D_1 cho đến D_i sẽ cho ra các đầu ra tương ứng từ t_1 cho đến t_i. Các kết quả đầu ra này là các cây con cho cây T.

Đây là 4 bước trong quá trình thiết kế và xây dựng cây quyết định [1].

Thuật toán:

Design_Tree (Bảng dữ liệu D, nút t, Chia_Chọn_Điều kiện C)

{

Thực hiện điều kiện C trên D để tìm ra các đầu ra có thể (t_1 đến t_i).

If (t không là nút lá)

Tạo ra nút trong của t và Chia D thành các tập dữ liệu con.

Thực hiện Đệ quy quá trình trên với các tập dữ liệu con D_i

EndIF

}

Tuyển nhân viên cho một công ty

Ví dụ: Giả sử chúng ta có 1 bảng dữ liệu về nhân sự của 1 công ty phần mềm đặt tên là bảng XYZ. Để tuyển mới những ứng viên (có thể không hoặc có kinh nghiệm), những người có thể đáp ứng được một số điều kiện nhất định, công ty cần phải lọc theo những thông tin mà ứng viên đã đăng ký [5]. Công ty yêu cầu những thông tin chi tiết từ ứng viên như tên, tên bố, tình trạng của ứng viên (đã có kinh nghiệm hay chưa), CPGA (một loại điểm tổng hợp) và tổng số năm kinh nghiệm đối với những ứng viên thi tuyển chức danh chuyên gia. Như thế bảng dữ liệu sẽ bao gồm những bản ghi sau:

Bảng 1. Các bản ghi của bảng dữ liệu

AID	Tên	Tên bố	Tình trạng ứng viên	Yêu cầu kinh nghiệm	CPGA
A01	Pam	Peter	Chưa có kinh nghiệm	0 năm	7.0
A02	Jin	Paul	Chưa có kinh nghiệm	0 năm	7.5
A03	Mick	Lee	Đã có kinh nghiệm	3 năm	6.0
A04	Nina	Pat	Đã có kinh nghiệm	3 năm	7.5
A05	Sam	Duke	Đã có kinh nghiệm	2 năm	7.5
A06	Leo	Mike	Đã có kinh nghiệm	2 năm	6.0

Bảng dữ liệu các nhân viên mới này bao gồm các bản ghi trên. Bây giờ vấn đề chính là chúng ta phải lựa chọn chỉ 1 số ít bản ghi phù hợp. Chính vì vậy chúng ta phải chuẩn bị một số điều kiện nhằm giảm số lượng bản ghi xuống và chọn được ứng viên thích hợp. Vấn đề chính là làm thế nào để mô hình hóa điều kiện này trong 1 cấu trúc cung cấp giải pháp cho vấn đề này với chỉ 1 trong 2 khả năng: đồng ý hay từ chối cho cuộc phỏng vấn.

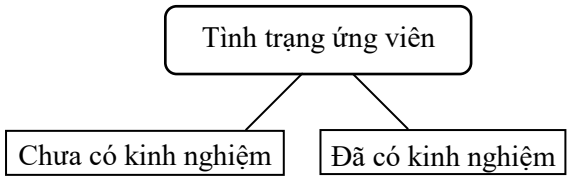
Đây sẽ là cây quyết định dưới dạng cây phân cấp quyết định [1] có thể là giải pháp cho vấn đề này. Chúng ta có thể sử dụng 4 bước thiết kế cây như ví dụ dưới đây:

Bước 1: Bảng dữ liệu D:

{AI01,AI02,...,AI06} có 6 bản ghi. Chúng ta sang bước 2:

Bước 2: Giả sử chúng ta sử dụng tình trạng ứng viên là điều kiện để kiểm tra với 2 đầu ra là <t_1,t_2> tương ứng với không có kinh nghiệm và đã có kinh nghiệm như

Hình 1. Cây quyết định với nút gốc và 2 đầu ra tương ứng "Chưa có kinh nghiệm" và" Đã có kinh nghiệm"

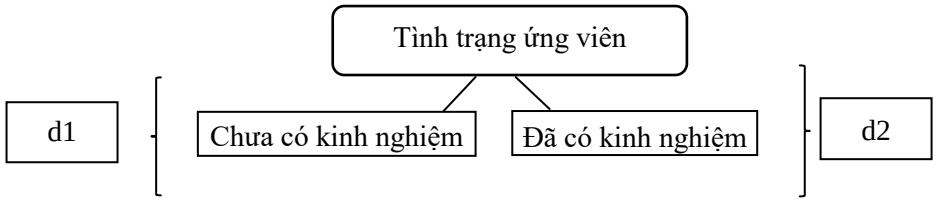


Bước 3: Bây giờ bảng dữ liệu D sẽ được chia thành 2 bảng dữ liệu con là D_1 và D_2 trong đó:

D_1:{AI01, AI02}

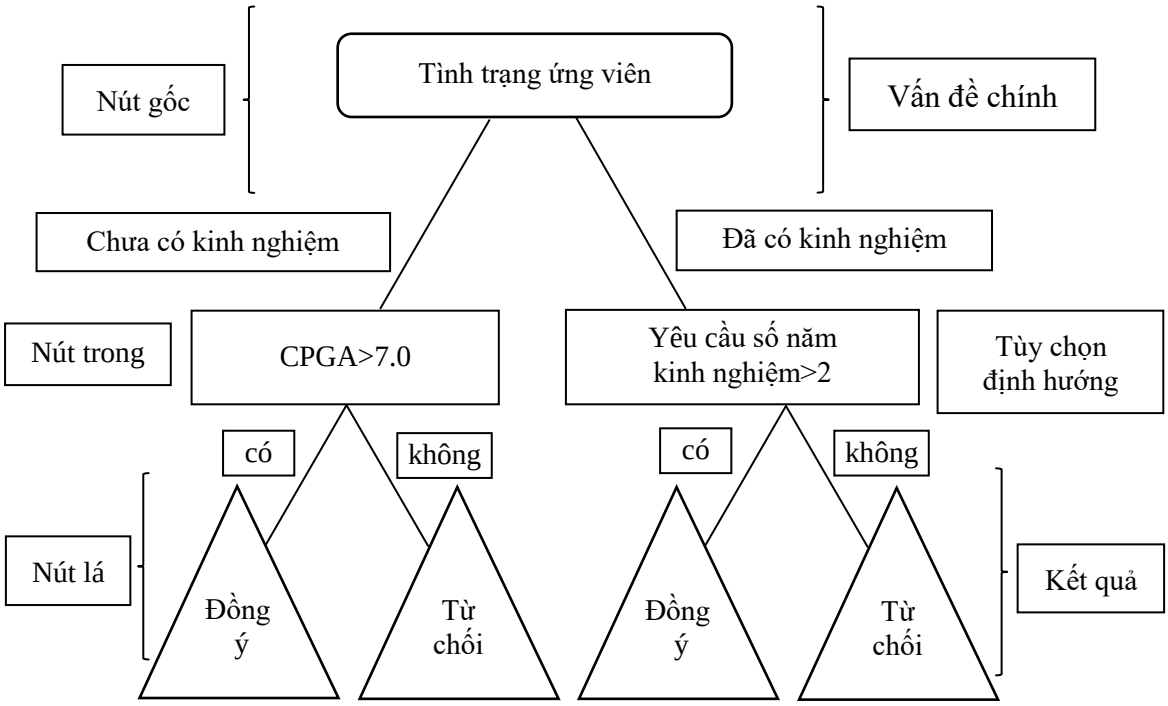
D_2:{AI03,AI04, AI05,AI06}

Hình 2. Cây quyết định có 2 bảng dữ liệu con $d_1$và d_2 và 2 đầu ra tương ứng.



Bước 4: Tiếp tục lặp lại quá trình này cho $d_1$và d_2 cho đến khi cây quyết định được xây dựng

Hình 3. Cây quyết định cuối cùng



Bây giờ cây quyết định cuối cùng cho bảng dữ liệu D là:

Bảng 2. Quyết định cuối cùng

AID	Quyết định
A01	Từ chối
A02	Đồng ý
A03	Từ chối
A04	Đồng ý
A05	Từ chối
A06	Từ chối

Từ ví dụ này chúng ta có thể hiểu làm thế nào mà cây quyết định có thể là 1 công cụ quan trọng để tìm ra được giải pháp cho 1 vấn đề.

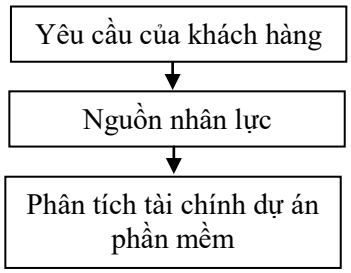
2.2. Các ứng dụng của cây quyết định

Phần này bao gồm các ứng dụng của cây quyết định trong các ứng dụng cụ thể và miêu tả làm thế nào các hiểu biết về cây quyết định có thể được sử dụng để giải quyết vấn đề.

2.2.1 Yêu cầu chấp thuận một dự án phần mềm

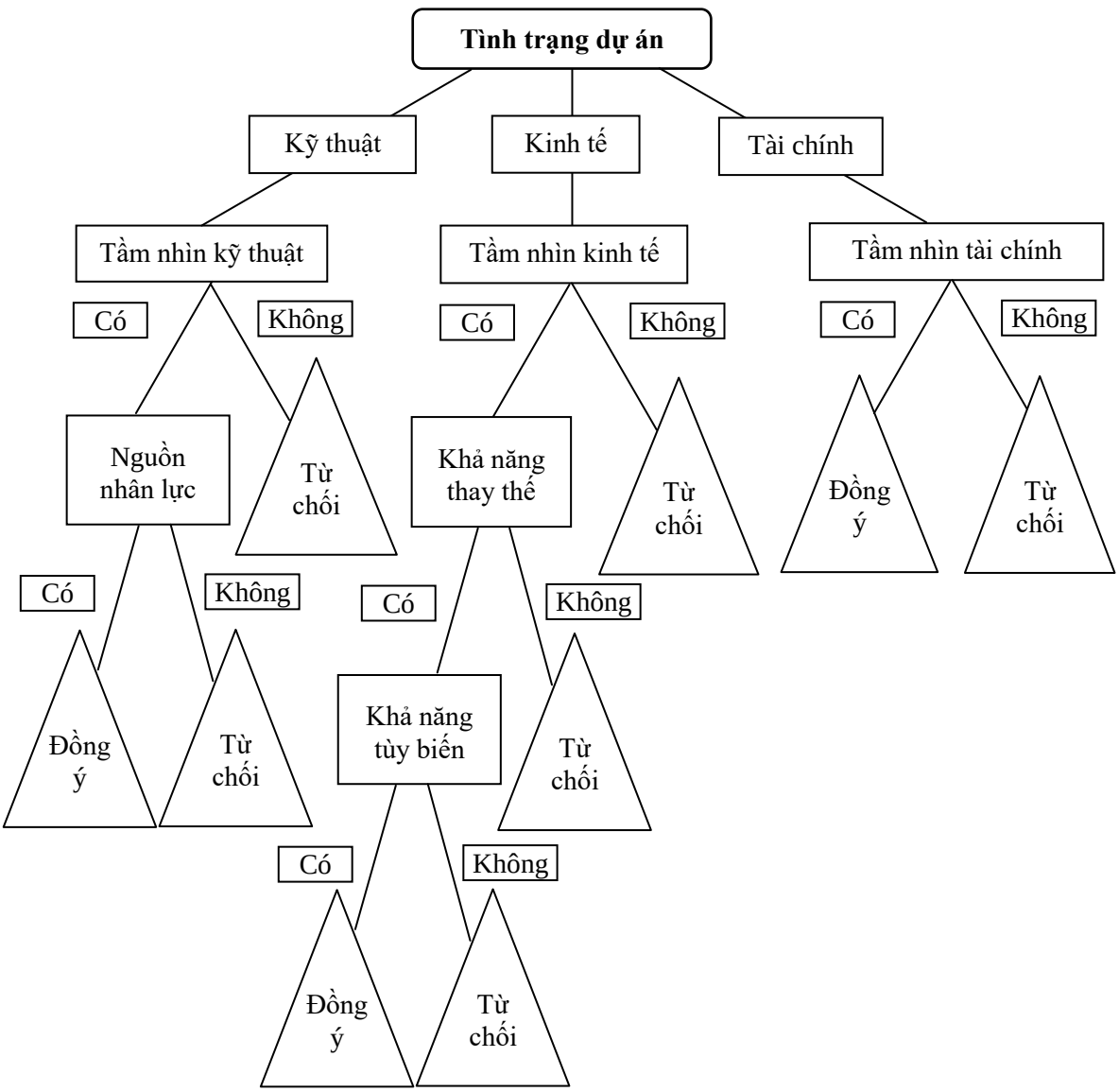
Trong ngành công nghiệp phần mềm khi 1 dự án đến với 1 công ty thì việc phân tích nguồn nhân lực và tài chính của công ty ấy sẽ đóng vai trò quan trọng trong việc có chấp thuận dự án ấy hay không [2]. Giả sử 1 khách hàng cần phần mềm có thể đáp ứng được các yêu cầu của họ. Từ đó công ty cần có nguồn nhân lực thích hợp cho việc sản xuất phần mềm đó, như là những chuyên gia viết phần mềm có kinh nghiệm. Nếu tất cả các yêu cầu đã được thỏa mãn thì chúng ta sẽ chuyển sang bước phân tích tài chính của dự án [5].

Hình 4. Quá trình thiết kế phần mềm



Bây giờ việc phân tích tài chính [5] sẽ chuẩn bị cây quyết định với kết quả chấp nhận hoặc từ chối yêu cầu thực hiện dự án. Nếu dự án được thông qua, giám đốc sẽ đưa ra các bước triển khai tiếp theo [3, 8], và sau đó quá trình phát triển phần mềm sẽ bắt đầu [4].

Hình 5.Yêu cầu chấp nhận dự án



2.2.2. Chính sách bảo hiểm

Cây quyết định [1] có thể được sử dụng trong dự đoán rủi ro khi đăng ký bảo hiểm cho 1 ứng viên trong đó rủi ro có thể xảy ra cho 1 ứng viên được đại diện là nút lá trong cây quyết định. Lấy ví dụ cụ thể sau:

Giả sử 1 công ty bảo hiểm bắt đầu 1 chính sách bảo hiểm có tên là PQR. Nếu có bất kỳ ứng viên nào muốn tham gia bảo hiểm thì ứng viên đó phải tuân theo 1 số điều kiện. Nếu có bất kỳ 1 điều kiện nào không được tuân theo thì có thể đó là sự mạo hiểm khi bảo hiểm cho ứng viên ấy. Do đó tất cả các yêu cầu phải hiện có trên cây quyết định mà từ đó có thể đưa ra 2 quyết định: Đăng ký hay từ chối cho ứng viên được tham gia bảo hiểm. Đăng ký nếu ứng viên cho thấy không có sự rủi ro nào từ các yêu cầu bảo hiểm và từ chối nếu chỉ có 1 rủi ro.

Để thực hiện chính sách bảo hiểm này cần có bảng thống kê cụ thể dựa theo các yếu tố có thể tác động lên quyết định mua bảo hiểm của người dân, bao gồm từ các đại lý bảo hiểm, quảng cáo họ nghe được hay từ lời khuyên của họ hàng người thân.

Bảng 3. Tổng hợp sự ảnh hưởng của các nhân tố đối với các quyết định mua bảo hiểm

Phân loại	Số lượng người trả lời	Phần trăm
Đại lý	77	64,17
Bạn thân hoặc họ hàng	15	12.50
Quảng cáo	6	5
Thành viên gia đình	9	7,50
Tự quyết định	13	10,83
Tổng	120	100

Ngoài ra còn phải tính đến các mục tiêu khi mua bảo hiểm của từng người. Điều này sẽ dẫn đến các yếu tố cấu thành chi tiết của 1 chính sách bảo hiểm riêng. Bảng thống kê sau được thực hiện dựa trên mục tiêu khi mua bảo hiểm của người dân. Từ bảng này ta có thể thấy đa phần người mua bảo hiểm cho gia đình mình là thứ yếu, do đó chúng ta sẽ nhấn mạnh yếu tố bảo vệ đối với chính sách bảo hiểm.

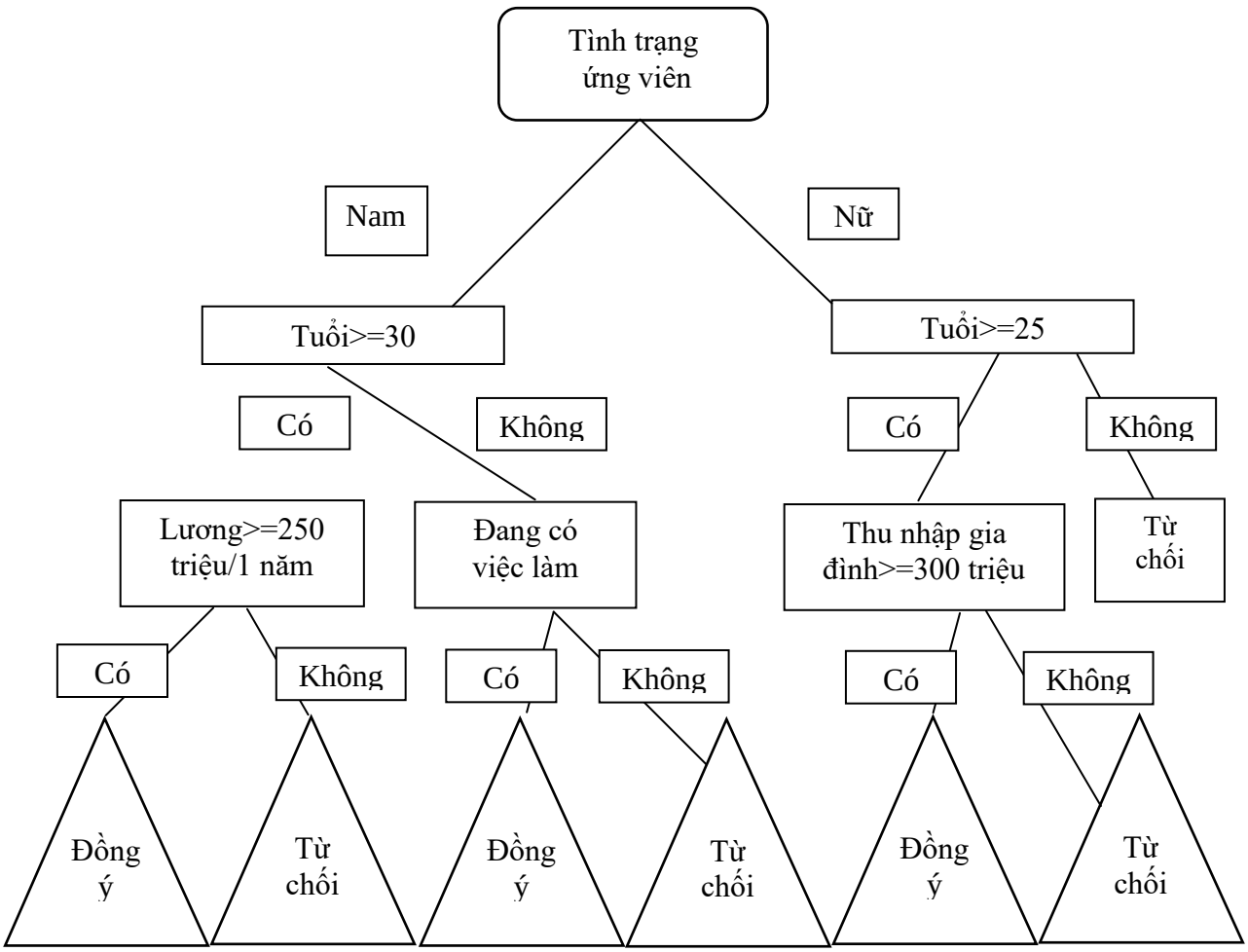
Bảng 4. Tổng hợp mục tiêu khi mua bảo hiểm

Mục tiêu khi mua bảo hiểm	Số lượng người trả lời	Phần trăm
Bảo vệ gia đình	100	83,34
Giảm thuế	10	8,33
Thu nhập tuổi già	10	8,33
Tổng	120	100

Ngoài ra còn vấn đề về tuổi tác khi có 1 số lượng người mua bảo hiểm như một sự đảm bảo tương lai cho tuổi già. Điều này dẫn tới việc chúng ta cho thêm yếu tố "tuổi" vào chính sách bảo hiểm của mình.Ví dụ đề xuất ở dưới với độ tuổi trên 30 đối với nam và trên 25 đối với nữ.

Từ các yếu tố trên ta có thể đưa ra cây quyết định với 1 ứng viên như sau:

Hình 6. Cây quyết định chấp thuận chính sách bảo hiểm



3. KẾT LUẬN

Bài báo đã thể hiện cách thiết kế cây quyết định cho 1 tình huống phức tạp như việc triển khai dự án phần mềm hay thực thi một chính sách bảo hiểm với các thống kê cụ thể. Tương lai tác giả sẽ đưa ra các thống kê cụ thể hơn tương ứng với những tình huống phức tạp hơn trong ứng dụng của cây phân cấp quyết định.

TÀI LIỆU THAM KHẢO

1. Robert N. Brticher (1999), *The limits of software: people, Project, and Perspectives*, Addison-Wesley Publisher.
2. Mark J Christensen, Richard H. Thayer(2002), *The Project Manager's Guide to Software Engineering's Best Practices*, Wiley-IEEE Computer Society.
3. N.E. Fenton (2014), *Software Metrics, A Rigorous & Practical Approach*, CRC Press.
4. Steve McConnell (2004), *Code Complete:A Practical Handbook of Software Construction*, Microsoft Press Publisher.
5. Dorothy Graham (2002), 'Requirements and Testing: Seven Missing-Link Myths', *IEEE Software*, volume: 19, pages:15-17.
6. Eric Fosler- Lussier (1999), *Multilevel decision trees for static and dynamic pronounciation models*, Uerospeech publisher.
7. Steve Pieczenik, Jeff Rovin (2006), *Divide and Conquer (Tom Clancy's Op-Center, Book 7)*, The Berkley Publishshing Group.
8. Gerard O'Regan (2002), *A Practical Approach to Software Quality Springer Verlag publisher*.

.