

ỨNG DỤNG KỸ THUẬT CÂY QUYẾT ĐỊNH TRONG KHAI PHÁ DỮ LIỆU XÂY DỰNG HỆ THỐNG TƯ VẤN CHỌN NGÀNH TUYỂN SINH ĐẠI HỌC

APPLYING DECISION TREE TECHNIQUE IN DATA MINING TO BUILD A CONSULTANT SYSTEM FOR CHOOSING MAJORS FOR UNIVERSITY ENTRANCE EXAMINATION

Nguyễn Văn Chức

Trường Đại học Kinh tế, Đại học Đà Nẵng; Email: chuc1803@gmail.com

Tóm tắt – Hiện nay, vấn đề tư vấn chọn ngành tuyển sinh đại học đang nhận được sự quan tâm rất lớn của xã hội. Mặc dù có rất nhiều websites tư vấn tuyển sinh, tuy nhiên các website này chỉ phục vụ cho việc tra cứu thông tin. Vấn đề cốt lõi của tư vấn tuyển sinh là làm sao giúp cho thí sinh có thể chọn được ngành học phù hợp với năng lực của mình. Bài báo này tập trung nghiên cứu kỹ thuật cây quyết định trong khai phá dữ liệu để xây dựng mô hình dự đoán nhằm tư vấn cho thí sinh có thể chọn được ngành học phù hợp với năng lực của mình. Dựa vào các tri thức phát hiện được từ mô hình dự đoán, một giao tiếp được xây dựng trên nền web để người dùng có thể dễ dàng sử dụng các tri thức này vào việc chọn ngành học cho mình.

Từ khóa – chọn ngành; cây quyết định; khai phá dữ liệu; mô hình dự đoán; tuyển sinh đại học.

1. Đặt vấn đề

Hiện nay, vấn đề tư vấn tuyển sinh đại học là nhu cầu cấp thiết đối với xã hội, nhất là các học sinh chuẩn bị dự thi đại học. Hàng năm, các trường đại học kết hợp với các cơ quan báo chí và các tổ chức xã hội tổ chức các đợt tư vấn tuyển sinh nhằm giúp cho thí sinh có được thông tin cần thiết để chọn ngành học phù hợp cho mình. Tuy nhiên, vấn đề cốt yếu của việc chọn ngành học phù hợp là người học cần phải hiểu rõ điểm mạnh của bản thân cũng như những yêu cầu để học tốt ngành học mà mình sẽ học. Bài báo này tập trung nghiên cứu về kỹ thuật phân lớp dữ liệu dựa vào cây quyết định trong khai phá dữ liệu để xây dựng mô hình dự đoán ngành học nhằm tư vấn cho thí sinh chọn ngành học phù hợp với năng lực của mình trên cơ sở nghiên cứu dữ liệu của sinh viên đang theo học các ngành kinh tế của trường Đại học Kinh tế - Đại học Đà Nẵng.

2. Giới thiệu về kỹ thuật phân lớp dữ liệu dựa vào cây quyết định

Trong lĩnh vực khai phá dữ liệu, cây quyết định (Decision Tree – DT) là một mô hình dự đoán (*predictive model*) thuộc lớp các bài toán phân lớp (classification problem) dùng để xác định lớp của các đối tượng cần dự đoán. Cây quyết định dựa vào dãy các luật để dự đoán lớp của đối tượng. Mỗi một nút trong (*internal node*) của DT tương ứng với một biến, đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá (*leaf*) đại diện cho giá trị dự đoán của biến phân loại. Cây quyết định học để dự đoán giá trị của các biến phân loại bằng cách dựa vào tập dữ liệu huấn luyện (*training data*) để chọn ra nút gốc (*root node*) để phân tách cây bằng cách tính độ lợi thông tin (Information Gain - IG), quá trình này được lặp lại một cách đệ quy cho đến khi không thể tiếp tục thực hiện

Abstract – Nowadays, society is interested in choosing majors for university entrance examination. Although there are a lot of websites of consultant university entrance examination, these websites are only used to search information. However how to help the candidates to decide the major of study consistent with their capabilities is the key this problem. This paper is focused on studying decision tree technique in data mining to build a predictive model which can be used to consult the candidates so they can choose major in line with their abilities. Based on the knowledge that was discovered from the predictive model, an interface is also built on a web platform to help users use this knowledge in choosing their majors of study.

Key words – choosing majors; decision tree; data mining; predictive model; university entrance examination.

việc phân tách cây được nữa. Kỹ thuật máy học (machine learning) dùng trong cây quyết định được gọi là học bằng cây quyết định và thường được gọi ngắn gọn là cây quyết định. [1], [2]

Dữ liệu huấn luyện cho cây quyết định là tập các bản ghi có dạng: $(x, y) = (x_1, x_2, \dots, x_k, y)$

Trong đó: y được gọi là biến phân loại (còn gọi là biến mục tiêu hay biến phụ thuộc) và $x_1, x_2, \dots, x_k$ là các biến độc lập.

Cây quyết định được chia thành hai loại:

Cây hồi quy (Regression Tree) dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu định lượng (quantitative) như dự đoán doanh thu, lợi nhuận, giá thành sản phẩm. . . . Thuật toán phổ biến dùng để xây dựng cây hồi quy là CART (Classification and Regression Trees).

Cây phân lớp (Classification Tree) dùng để dự đoán giá trị của biến phân loại có kiểu định danh (nominal) như dự đoán khả năng mua hàng của khách hàng (có mua hoặc không mua), khả năng bị bệnh của bệnh nhân (có bệnh hoặc không có bệnh), kết quả học tập của sinh viên (xuất sắc, giỏi, khá, trung bình, yếu). . . . Thuật toán phổ biến dùng để xây dựng cây phân lớp là ID3, C4.5. Trong đó thuật toán C4.5 được cải tiến từ thuật toán ID3.

Thuật toán ID3 xây dựng cây quyết định

Thuật toán ID3 (Iterative Dichotomiser 3)

Thuật toán ID3 do Ross Quinlan đề xuất, tư tưởng của thuật toán ID3 là việc xây dựng cây quyết định được thực hiện đệ quy từ trên xuống và sử dụng độ lợi thông tin (IG – Information Gain) làm độ đo để chọn node gốc để phân tách cây. IG là tham số được tính toán dựa trên Entropy trong lý thuyết thông tin. Node được chọn làm node gốc là node có IG lớn nhất (hoặc node có Entropy nhỏ nhất).

[2] Công thức tính Entropy và IG:

$$\text{Entropy}(S) = \sum_{I \in C} -p(I) \log_2 p(I)$$

Trong đó:

S: tập dữ liệu huấn luyện

$p(I)$: tỷ số giữa các mẫu thuộc về lớp I trên tổng số các mẫu huấn luyện trong S

C: tập các giá trị của thuộc tính phân loại

Công thức tính giá trị IG cho thuộc tính A:

$$\text{IG}(S, A) = \text{Entropy}(S) - \sum_{v \in A} ((|S_v|/|S|) * \text{Entropy}(S_v))$$

Trong đó:

- v: các giá trị của thuộc tính A
- S_v : tập con của tập S với các mẫu thuộc tính A có giá trị v
- $|S_v|$: số các mẫu thuộc S_v
- |S|: số mẫu của tập S

Các bước chính trong thuật toán ID3

- Tính Entropy của tất cả các thuộc tính trong data set S
- Chia tập S thành các tập con (subsets) sử dụng thuộc tính có Entropy nhỏ nhất (tương đương với IG lớn nhất).
- Tạo cây quyết định với node gốc là nút có IG lớn nhất
- Thực hiện đệ quy trên các subsets sử dụng các thuộc tính còn lại

ID3 Algorithm

ID3(R, C, S):

Input:

- R: tập các thuộc tính,
- C: thuộc tính phân loại
- S: tập dữ liệu huấn luyện

Output:

Returns a Decision Tree

Begin

If S rỗng **then** trả về một node với giá trị lỗi **Failure**;

If S gồm các records với giá trị thuộc tính phân loại giống nhau **then** trả về node với giá trị đó;

If R rỗng **then** trả về một node với giá trị có tần suất xuất hiện nhiều nhất trong các giá trị của thuộc tính phân loại trong S;

Let D là thuộc tính có giá trị **IG(S,D)** lớn nhất trong R;

Let $\{d_j | j = 1, 2, \dots, m\}$ là các giá trị của thuộc tính D;

Let $\{S_j | j = 1, 2, \dots, m\}$ là các tập con của S gồm các records tương ứng với giá trị d_j của thuộc tính D;

Return một cây với node gốc (root) có nhãn là D và các cạnh có nhãn $d_1, d_2, \dots, d_m$ tương ứng với các giá trị của thuộc tính D;

ID3(R-D, C, S₁), ID3(R-D, C, S₂), ...

ID3(R-D, C, S_m);

End ID3;

3. Ứng dụng cây quyết định xây dựng hệ thống tư vấn chọn ngành tuyển sinh đại học

3.1. Mô tả ứng dụng

Mục đích của ứng dụng: Nghiên cứu kỹ thuật phân lớp dữ liệu dựa vào cây quyết định để xây dựng mô hình dự đoán

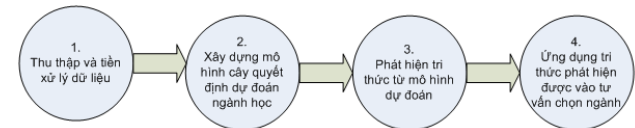
ngành học. Từ kết quả của mô hình dự đoán, một giao tiếp trên nền web được xây dựng để thí sinh có thể chọn ngành học phù hợp với năng lực của mình.

Dữ liệu đầu vào: Gồm các đặc trưng liên quan đến việc chọn ngành học của thí sinh như khả năng toán học, vật lý, hóa học, văn học, ngoại ngữ, năng khiếu, khả năng thuyết trình...

Đầu ra: Ngành học phù hợp nhất với khả năng của thí sinh.

3.2. Quy trình triển khai hệ thống tư vấn chọn ngành

Quy trình triển khai hệ thống tư vấn chọn ngành tuyển sinh được tiến hành theo các bước chính như Hình 1.



Hình 1: Quy trình triển khai hệ thống tư vấn chọn ngành

Bước 1: Khảo sát và thu thập dữ liệu:

Dữ liệu dùng để xây dựng cây quyết định chọn ngành tuyển sinh được thu thập thông qua bảng hỏi để thu thập dữ liệu về các yếu tố ảnh hưởng đến việc chọn ngành của thí sinh như khả năng học các môn tự nhiên, các môn xã hội, khả năng ngoại ngữ, năng khiếu của thí sinh và một số yếu tố khác như khả năng thuyết trình, kỹ năng lãnh đạo nhóm, hoàn cảnh kinh tế... Dữ liệu điều tra ban đầu để xây dựng mô hình dự đoán ngành học gồm rất nhiều thuộc tính, sau quá trình tiền xử lý dữ liệu (sử dụng phương pháp trích chọn thuộc tính) để đánh giá mức độ ảnh hưởng của các thuộc tính đến việc chọn ngành, mô hình xác định được các thuộc tính có ảnh hưởng đến việc chọn ngành như Bảng 1.

Dữ liệu thu thập gần 1500 mẫu từ các sinh viên năm 1 của trường Đại học Kinh tế - Đại học Đà Nẵng theo cấu trúc sau của Bảng 1.

Bước 2. Xây dựng mô hình cây quyết định dự đoán ngành học

Mô hình cây quyết định dự đoán ngành được xây dựng trên công cụ khai phá dữ liệu Business Intelligence Development Studio (BIDS) của Microsoft. BIDS là công cụ rất mạnh cho phép triển khai các mô hình khai phá dữ liệu, được sử dụng rộng rãi hiện nay bởi khả năng kết nối dễ dàng với nhiều nguồn dữ liệu, giao diện dễ sử dụng và nhất là khả năng biểu diễn tri thức phát hiện được rất trực quan dễ hiểu, dễ sử dụng. BIDS được tích hợp vào SQL SERVER 2005 trở về sau trong các phiên bản Enterprise hoặc Development. [3]

Sau khi thực hiện các thao tác tiền xử lý dữ liệu để phù hợp với mô hình khai phá dữ liệu, sử dụng Microsoft Decision Tree trong BIDS để xây dựng cây quyết định chọn ngành. Kết quả cây quyết định dự đoán ngành như Hình 5.

Bước 3. Phát hiện tri thức từ mô hình cây quyết định

Từ cây quyết định đã xây dựng, các tri thức phát hiện được cho dưới dạng các luật:

IF L_1 AND L_2 AND ... AND L_n THEN Ngành = "M".

Trong đó:

Bảng 1: Cấu trúc của training data tư vấn chọn ngành

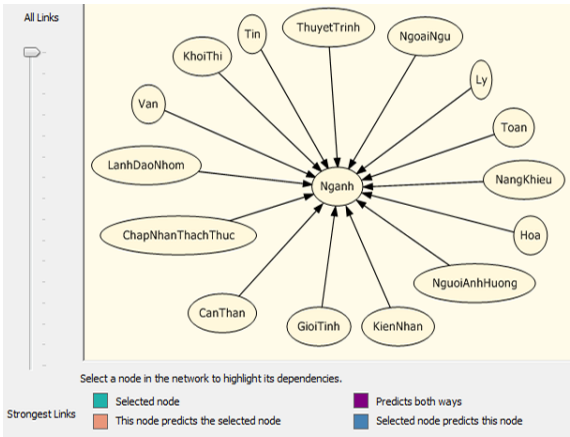
STT	Thuộc tính	Kiểu DL	Giá trị của thuộc tính	Giải thích
1	GioiTinh	Nominal	Nam, Nữ	Giới tính
2	TinhTp	Nominal	Các tỉnh/Thành phố	Tỉnh/Thành phố
3	KhoiThi	Nominal	A,A1,D1,D2, D3, D4	Khối thi
4	NangLucToan	Nominal	Xuất sắc, Giỏi, Khá, Trung Bình, Yếu	Năng lực Toán học
5	NangLucLy	Nominal	Xuất sắc, Giỏi, Khá, Trung Bình, Yếu	Năng lực Vật lý
6	NangLucToan	Nominal	Xuất sắc, Giỏi, Khá, Trung Bình, Yếu	Năng lực Hóa học
7	NangLucVan	Nominal	Xuất sắc, Giỏi, Khá, Trung Bình, Yếu	Năng lực Văn học
8	NangLucNgoaiNgu	Nominal	Xuất sắc, Giỏi, Khá, Trung Bình, Yếu	Năng lực Ngoại ngữ
9	NangLucTin	Nominal	Xuất sắc, Giỏi, Khá, Trung Bình, Yếu	Năng lực Tin học
10	ThuyetTrinh	Nominal	Rất tốt, Tốt, Bình thường, Không tốt, Rất không tốt	Khả năng Thuyết trình
11	KienNhan	Nominal	Rất kiên nhẫn, Kiên nhẫn, Bình thường, Ít kiên nhẫn, Không kiên nhẫn	Tính Kiên nhẫn
12	CanThan	Nominal	Rất cẩn thận, Cẩn thận, Bình thường, Ít cẩn thận, Không cẩn thận	Tính Cẩn thận
13	SangTao	Nominal	Rất sáng tạo, Sáng tạo, Bình thường, Ít sáng tạo, Không sáng tạo	Khả năng Sáng tạo
14	LanhDaoNhom	Nominal	Rất tốt, Tốt, Bình thường, Không tốt, Rất không tốt	Khả năng Lãnh đạo nhóm
15	ChapNhanThachThuc	Nominal	Rất tốt, Tốt, Bình thường, Không tốt, Rất không tốt	Khả năng chấp nhận thách thức trong công việc
16	NangKhieu	Nominal	Âm nhạc, Điện ảnh, Hội họa, Thể thao,	Không có Năng khiếu
17	NguoiAnhHuong	Nominal	Ba mẹ, Anh chị em, Bạn bè, Bản thân, Thầy cô giáo, Khác	Người ảnh hưởng trong việc chọn ngành
18	DieuKienKTGiaDinh	Nominal	Rất cao, Cao, Trung bình, Thấp, Rất thấp	Điều kiện Kinh tế Gia đình
19	NganhHoc	Nominal	Các ngành học	Ngành học (thuộc tính phân loại)

- $L_1, L_2, \dots, L_n$: là các biểu thức logic;
 - M: là một ngành học cụ thể nào đó mà về trái là các thuộc tính và về phải là giá trị có thể có của thuộc tính đó.
 Chẳng hạn, hai luật được trích ra từ cây quyết định chọn ngành đã xây dựng như sau:

Luật 1: IF $Ly = \text{"Giỏi"}$ and $NangKhieu = \text{"Thể thao"}$ and $NgoaiNgu = \text{"Trung Bình"}$ and $GioiTinh = \text{"Nam"}$ THEN $Nganh = \text{"Kiểm toán"}$

Luật 2: IF $Ly = \text{"Giỏi"}$ and $NangKhieu = \text{"thể thao"}$ and $NgoaiNgu = \text{"Trung Bình"}$ and $GioiTinh = \text{"Nữ"}$ THEN $Nganh = \text{"Kế toán"}$

Ngoài ra, mạng phụ thuộc của mô hình cho biết độ mạnh (weight) của các nhân tố ảnh hưởng đến việc chọn ngành.



Hình 2: Mạng phụ thuộc của mô hình

Dựa vào các tri thức phát hiện được từ mô hình cây quyết định dự đoán ngành học đã xây dựng, một hệ thống giao tiếp được xây dựng trên nền web cho phép người dùng có thể sử

dụng các tri thức này vào việc chọn ngành học cho mình bằng cách cung cấp các thông tin liên quan đến việc chọn ngành được sử dụng trong mô hình. Hệ thống sẽ đề xuất cho người dùng lựa chọn ngành học phù hợp với các thông tin mà người dùng đã cung cấp.

TƯ VẤN CHỌN NGÀNH ĐĂNG KÝ TUYỂN SINH

[Loc thông tin](#)

Tỉnh thành

DaNang

Tỉnh / Thành phố bạn ở

Nu

Giới tính

A

Khối thi dự kiến

Gioi

Học lực môn Toán

Kha

Học lực môn Lý

Kha

Học lực môn Hóa

Trung Binh

Học lực môn Ngoại ngữ

Trung Binh

Học lực môn Văn

Kha

Học lực môn Tin học

Tot

Khả năng thuyết trình

BinhThuong

Kiên nhẫn

RaiCanThan

Cẩn thận

ItSangTao

Sáng tạo

KhongTot

Khả năng lãnh đạo

BinhThuong

Khả năng vượt qua thách thức

HoiHoa

Năng khiếu

BeMe

Người ảnh hưởng

Thap

Điều kiện kinh tế

DuDoan

Reset

Hình 3: Giao tiếp người dùng với hệ thống tư vấn chọn ngành

TƯ VẤN CHỌN NGÀNH ĐĂNG KÝ TUYỂN SINH

[Loc thông tin](#)

Kết quả 3 ngành phù hợp nhất với bạn (xếp theo mức độ phù hợp giảm dần):

Ngành	Mức độ phù hợp (%)
KeToan	80,0
KinhTePhatTrien	10,0
QTKDTongQuat	10,0

Hình 4: Kết quả dự đoán ngành từ mô hình

4. Kết luận và hướng phát triển

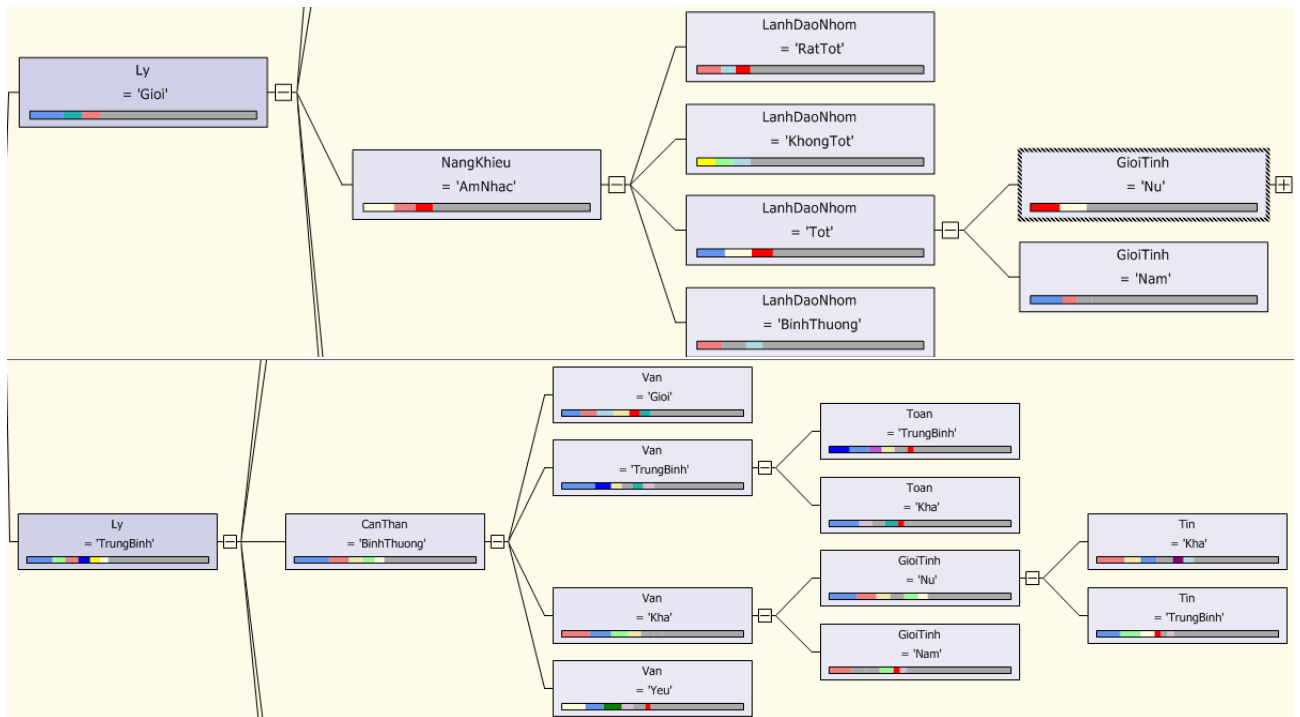
Khai phá dữ liệu ngày càng được sử dụng rộng rãi trong quá trình phát hiện tri thức trên khối lượng dữ liệu lớn nhằm hỗ trợ ra quyết định. Cây quyết định là kỹ thuật được sử

dụng phổ biến để giải quyết bài toán phân lớp dữ liệu bởi tính đơn giản, hiệu quả và nhất là khả năng biểu diễn tri thức phát hiện được rất trực quan, dễ hiểu, dễ sử dụng. Bài báo đã tìm hiểu về lý thuyết cây quyết định, từ đó nghiên cứu ứng dụng kỹ thuật này vào xây dựng mô hình dự đoán ngành học. Trên cơ sở các tri thức phát hiện được từ mô hình cây quyết định đã xây dựng, một giao tiếp được xây dựng trên nền web giúp cho người dùng có thể dễ dàng sử dụng các tri thức này vào việc chọn ngành phù hợp với năng lực của mình bằng cách cung cấp nhưng thông tin liên quan đến việc dự đoán ngành học mà mô hình sử dụng.

Hiện nay, có rất nhiều hệ thống (website) tư vấn tuyển sinh trực tuyến. Tuy nhiên, các hệ thống này chỉ dừng lại ở việc cho phép tra cứu thông tin liên quan đến tuyển sinh

như số hồ sơ, tỷ lệ chọi, điểm chuẩn... của các trường tuyển sinh chứ chưa giải quyết được vấn đề quan trọng của công tác tư vấn tuyển sinh đó là tư vấn chọn ngành học. Kết quả nghiên cứu của bài báo có thể tích hợp vào các hệ thống tư vấn tuyển sinh hiện có để nâng cao hiệu quả của công tác tuyển sinh trực tuyến.

Hạn chế của mô hình là do dữ liệu thu thập chưa thật đầy đủ, dữ liệu huấn luyện mô hình được thu thập từ dữ liệu của sinh viên trường Đại học Kinh tế - Đại học Đà Nẵng. Vì vậy, mô hình chỉ dự đoán và tư vấn cho các ngành thuộc khối ngành kinh tế. Trong thời gian tới, sẽ tiếp tục thu thập thêm dữ liệu để hoàn thiện mô hình, đồng thời nghiên cứu phát triển mô hình để có thể dự đoán và tư vấn chọn ngành cho các khối ngành khác như kỹ thuật, sư phạm, xã hội...



Hình 5: Một nhánh của cây quyết định chọn ngành

Tài liệu tham khảo

- [1] David Squire, CSE5230 Tutorial: The ID3 Decision Tree Algorithm, Faculty of Information Technology, Monash University, 2004.
- [2] Rokach Lior; Maimon O., Data mining with decision trees: theory and applications, World Scientific Pub Co Inc. ISBN 978-9812771711, 2008.
- [3] Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat, Data Mining with Microsoft SQL Server 2008, ISBN: 978-0-470-27774-4, 2008.
- [4] <http://www.sqlserverdatamining.com>.
- [5] <http://msdn.microsoft.com/en-us/library/ms173767.aspx> (Introducing Business Intelligence Development Studio)
- [6] <http://bis.net.vn/forums/t/458.aspx> (Giới thiệu công cụ xây dựng mô hình khai phá dữ liệu Business Intelligence Development Studio của Microsoft)

(BBT nhận bài: 12/12/2013, phản biện xong: 25/12/2013)