

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN XUÂN HẢI

**KHAI PHÁ DỮ LIỆU
VÀ ỨNG DỤNG TRONG DỰ BÁO TIẾN TRÌNH HỌC TẬP
CỦA SINH VIÊN ĐẠI HỌC THỦY LỢI**

CHUYÊN NGÀNH : KHOA HỌC MÁY TÍNH

MÃ SỐ: 60.48.01.01

LUẬN VĂN THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN ĐÌNH HÓA

HÀ NỘI - 2016

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Nguyễn Đình Hóa

Phản biện 1: TS. Phạm Văn Cường

Phản biện 2: TS. Tạ Quang Hùng

Luận văn được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 8 giờ 40 ngày 20 tháng 08 năm 2016

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Tính cấp thiết của đề tài

Mục tiêu chung của các em sinh viên cũng như của các bậc phụ huynh khi bước chân vào Trường Đại học chính là tấm bằng Đại học. Tuy nhiên, với mô hình đào tạo theo tín chỉ hiện nay tại hầu hết các trường Đại học nói chung và Đại học Thủy lợi nói riêng thì việc tìm hiểu, thích nghi với quy chế đào tạo mới là một điều không hề dễ dàng (trong quá trình học 12 năm phổ thông thì người học được đào tạo theo niên chế). Trong thực tế, rất nhiều sinh viên vẫn giữ thói quen cũ từ thời phổ thông (thang điểm, phương thức học tập...) trong quá trình học đại học, từ đó phát sinh ra những trường hợp đáng tiếc mà do thiếu hiểu biết, các em đã bỏ lỡ mất cơ hội của mình. Ví dụ như tại Đại học Thủy lợi, có trường hợp sinh viên học đạt hầu hết các môn (Điểm D tức là từ 4-5.4 điểm theo thang 10 là đạt [1]) nhưng lại không đủ điều kiện làm Đồ án tốt nghiệp (điều kiện làm Đồ án tốt nghiệp là không nợ môn và điểm trung bình chung các môn là 2.0 theo thang điểm 4 [3]), từ đó dẫn đến việc em bị chậm tiến độ học tập...

Để các em sinh viên và phụ huynh phần nào có cái nhìn rõ ràng hơn về tương lai việc học tập tại Trường Đại học mà không cần phải hiểu sâu về quy chế đào tạo theo tín chỉ: Đó là khả năng hoàn thành chương trình học như thế nào? Có đảm bảo tiến độ theo khung chung của nhà trường hay không? Có nguy cơ bị cảnh báo học tập hay không?... Từ đó, các em và gia đình có thể có những quyết định hợp lý, kịp thời trong thời gian học tập. Giải pháp tác giả đưa ra là cung cấp cho sinh viên và gia đình thông tin dự báo về tiến trình học tập trong tương lai của sinh viên dựa trên những dữ liệu hiện tại của sinh viên. Thông qua đó, sinh viên sẽ có thể đưa ra được những quyết định kịp thời, hợp lý cho việc học tập của mình; nhà trường cũng có thể có những giải pháp kịp thời để quan tâm, cảnh báo, khuyến khích các em sinh viên; gia đình cũng có thể nhìn nhận và hỗ trợ, động viên con em của mình...

Xuất phát từ thực tế và mục tiêu như vậy, tác giả thực hiện đề tài luận văn có tên “Khai phá dữ liệu và ứng dụng trong dự báo tiến trình học tập của sinh viên Đại học Thủy lợi” để giải quyết vấn đề nêu trên.

Tổng quan về vấn đề nghiên cứu

Khai phá dữ liệu (data mining) là quá trình khám phá các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có. Khai phá dữ liệu đã và đang được ứng dụng rộng rãi trong rất nhiều lĩnh vực hiện nay như: Tài chính, chứng khoán; Sinh học; Viễn thông...

Dự báo là tiên đoán những sự việc sẽ xảy ra trong tương lai, trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được; nói cách khác, dự báo được rút ra từ mô hình được xây dựng từ các đặc trưng dữ liệu được trích xuất ra từ bộ dữ liệu ban đầu sau khi khai phá dữ liệu. Trong thời đại công nghệ thông tin và toàn cầu hóa, dự báo đóng vai trò ngày càng quan trọng khi nhu cầu về thông tin tại thời điểm nào đó trong tương lai ngày càng lớn. Trong thực tế, có rất nhiều các mô hình dự báo được ứng dụng trong rất nhiều lĩnh vực thực tế, ví dụ như dự báo khí tượng thủy văn (sử dụng mô hình GSM, HRM...), dự báo tỷ giá hay chứng khoán (sử dụng mô hình ARIMA), dự báo về sử dụng điện năng (mô hình mạng nơron...), hay trong giáo dục, gần đây có nghiên cứu về dự báo kết quả thi đại học từ kết quả thi đại học và dữ liệu điểm các môn học sẽ thi đại học từ 03 năm học phổ thông (sử dụng Cây quyết định, K láng giềng gần nhất).

Tuy nhiên, hiện vẫn chưa có nghiên cứu cụ thể nào có thể giải quyết bài toán thực tế mà đề tài luận văn nhắc đến ở trên. Do đó, tác giả tiến hành thực hiện đề tài luận văn nghiên cứu về vấn đề khai phá dữ liệu và ứng dụng vào giải quyết bài toán thực tế là dự đoán tiến trình học tập của sinh viên Đại học Thủy lợi

Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu

Luận văn tiến hành nghiên cứu, tìm hiểu các vấn đề cơ bản về khai phá dữ liệu, các công cụ học máy. Từ đó ứng dụng vào việc xây dựng mô hình dự báo tiến trình học tập của sinh viên Đại học Thủy lợi. Qua luận văn này, tác giả mong muốn có những nghiên cứu lý thuyết về khai phá dữ liệu, các công cụ học máy và các thuật toán dự báo (Cây quyết định, K láng giềng gần nhất); thực nghiệm, phân tích được kết quả dự báo tiến trình học tập của sinh viên.

Thông qua phương pháp nghiên cứu lý thuyết và phương pháp nghiên cứu thực nghiệm, tác giả đã tiếp cận nghiên cứu các văn bản pháp quy và các hướng dẫn thực hiện quy chế đào tạo theo tín chỉ; dữ liệu về chương trình đào tạo tạo, điểm, kết quả học vụ của sinh viên Đại học Thủy lợi hay các công nghệ liên quan đến khai phá dữ liệu để tổng hợp thu thập thông tin. Từ đó phân tích được các yêu cầu của công việc, vận dụng các kết quả lý thuyết vào bộ dữ liệu cụ thể của Trường Đại học Thủy lợi để đánh giá và phân tích kết quả

Cấu trúc luận văn

Nội dung của luận văn được trình bày trong ba phần chính như sau:

1. Phần mở đầu
2. Phần nội dung: bao gồm ba chương

Chương 1: Tổng quan về Khai phá dữ liệu trong bài toán dự báo

Chương 2: Khai phá dữ liệu và các công cụ học máy

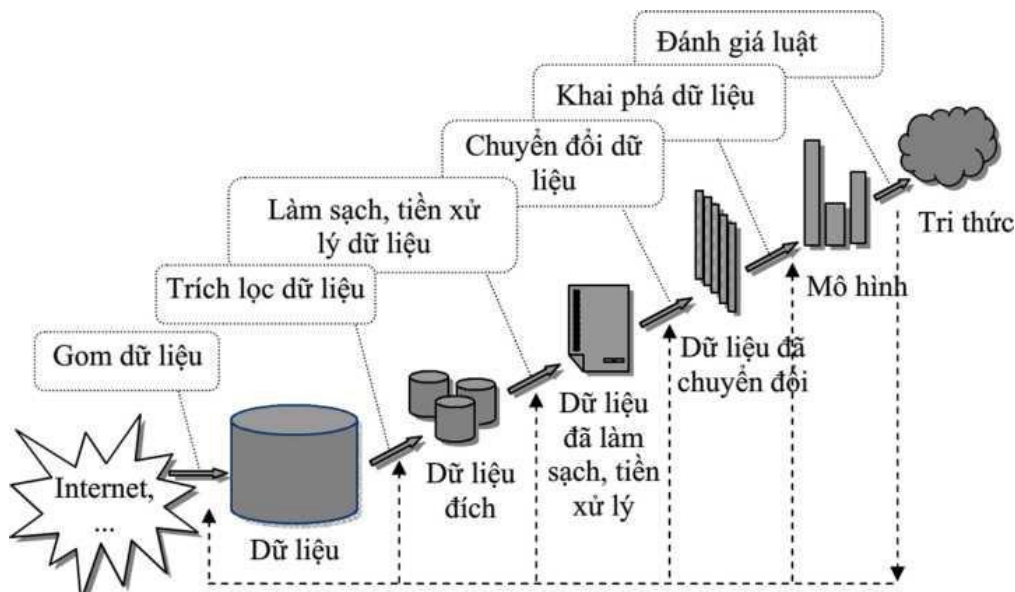
Chương 3: Dự báo tiến trình học tập của sinh viên Đại học Thủy lợi

3. Phân kết luận

CHƯƠNG I. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU TRONG BÀI TOÁN DỰ BÁO

1.1 Tổng quan về khai phá dữ liệu

Khai phá dữ liệu (data mining) là quá trình khám phá các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có. Quy trình khám phá tri thức trong cơ sở dữ liệu (KDD - Knowledge Discovery in Databases) thường tuân theo các bước như hình 1.1 dưới đây:



Hình 1.1. Quá trình khám phá, phát hiện tri thức từ dữ liệu [4]

1.2 Một số phương pháp khai phá dữ liệu

1.2.1 Phân lớp (Classification)

1.2.2 Phân cụm (Clustering)

1.2.3 Luật kết hợp (Association Rules)

1.3 Tổng quan về bài toán dự báo

1.3.1 Khái niệm cơ bản

Dự báo (hay còn gọi là dự đoán, tiên lượng) là tiên đoán những sự việc sẽ xảy ra trong tương lai, dựa trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được; nói cách khác, dự báo được rút ra từ mô hình được xây dựng từ các đặc trưng dữ liệu được trích xuất ra từ bộ dữ liệu ban đầu sau khi khai phá dữ liệu.

Dự báo dữ liệu là một quá trình gồm hai bước, nó gần giống với quá trình phân lớp. Tuy nhiên để dự đoán, chúng ta bỏ qua khái niệm nhãn phân lớp bởi vì các giá trị được dự đoán là liên tục (được sắp xếp) hơn là các giá trị phân loại. Ví dụ thay vì phân loại xem một khoản vay có là an toàn hay rủi ro thì chúng ta sẽ dự đoán xem tổng số tiền cho vay của một khoản vay là bao nhiêu thì khoản vay đó là an toàn. Do đó, ta có thể thấy rằng tất cả những đặc điểm của bài toán phân lớp hiện hữu trực tiếp tại bài toán dự báo

1.3.2 Đặc điểm của bài toán dự báo

Quá trình dự báo thường gồm 2 bước:

Bước 1: Xây dựng mô hình

Trong bước này, một mô hình sẽ được xây dựng dựa trên việc phân tích các mẫu dữ liệu sẵn có. Đây là quá trình học, trong đó một thuật toán phân lớp được xây dựng bằng cách phân tích hoặc “học” từ tập dữ liệu huấn luyện được xây dựng sẵn bao gồm nhiều bộ dữ liệu (xem ví dụ ở Hình 1.2).

Bước 2: Sử dụng mô hình đã xây dựng để phân lớp, dự báo dữ liệu

Trong bước này mô hình thu được sẽ được sử dụng để phân lớp, dự báo. Việc đầu tiên trong bước này là phải làm là tính độ chính xác của mô hình. Để đảm bảo tính khách quan nên áp dụng mô hình này trên một tập kiểm thử hơn là làm trên tập dữ liệu huấn luyện ban đầu. Tính chính xác của mô hình phân lớp trên tập dữ liệu kiểm thử là số phần trăm các bộ dữ liệu kiểm tra được đánh nhãn đúng bằng cách so sánh chúng với các mẫu trong bộ dữ liệu huấn luyện. Nếu như độ chính xác của mô hình dự đoán là chấp nhận được thì chúng ta có thể sử dụng mô hình để dự đoán nhãn lớp cho các mẫu dữ liệu khác với thông tin nhãn phân lớp chưa xác định trong tương lai [8].

1.3.3 Các phương pháp đánh giá cho bài toán phân lớp, dự báo

Có nhiều kỹ thuật để có thể đánh giá độ chính xác của các thuật toán phân lớp. Trong đó Holdout và K-fold cross validation (đánh giá chéo dựa trên k phần) là hai kỹ thuật phổ biến để đánh giá độ chính xác phân lớp dựa trên các phân chia lấy mẫu ngẫu nhiên từ dữ liệu cho trước [8].

1.4 Một số kỹ thuật khai phá dữ liệu trong bài toán dự báo/phân lớp

1.4.1 Các phương pháp cây quyết định

Cây quyết định (Decision Tree) là cấu trúc cây có dạng biểu đồ luồng, mỗi nút trong là kiểm định trên một thuộc tính, mỗi nhánh đại diện cho một kết quả kiểm định, các nút lá đại diện cho các lớp. Nút cao nhất trên cây là nút gốc.

1.4.2 Các phương pháp K-láng giềng gần nhất

Ý tưởng thuật toán học K-láng giềng gần là “thực hiện như các láng giềng gần của bạn đã làm”. Để dự đoán hoạt động của một mẫu xác định, K-láng giềng tốt nhất của mẫu đó sẽ được xem xét, và trung bình các hoạt động của các láng giềng gần sẽ đưa ra được dự đoán về hoạt động của mẫu đó

1.4.3 Các phương pháp dựa trên luật

Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các thành phần dữ liệu trong cơ sở dữ liệu. Mẫu đầu ra của giải thuật khai phá dữ liệu là tập luật kết hợp tìm được. Một ví dụ đơn giản về luật kết hợp là sự kết hợp giữa hai thành phần A và B có nghĩa là sự xuất hiện của A trong bản ghi kéo theo sự xuất hiện của B trong cùng bản ghi đó: $A \Rightarrow B$.

1.4.4 Các phương pháp Bayes «ngây thơ» và mạng tin cậy Bayes

Phân lớp Bayesian là phân lớp thống kê. Phân lớp Bayesian dựa trên định lý Bayes. Một phân lớp đơn giản của Bayesian đó là Naive Bayesian, so với việc thực thi của phân lớp cây quyết định và mạng nơron, phân lớp Bayesian đưa ra độ chính xác cao và nhanh khi áp dụng vào các cơ sở dữ liệu lớn.

1.5 Kết luận chương 1

CHƯƠNG II. KHAI PHÁ DỮ LIỆU VÀ CÁC CÔNG CỤ HỌC MÁY

1.1 Cây quyết định

1.1.1 Tổng quan về cây quyết định

1.1.1.1 Giới thiệu chung

Cây quyết định (decision tree) là một phương pháp mạnh và thường được sử dụng cho cả hai nhiệm vụ của khai phá dữ liệu là phân loại và dự báo. Mặt khác, cây quyết định còn có thể chuyển sang dạng biểu diễn tương đương dưới dạng tri thức với các luật If-Then. Cây quyết định là cấu trúc biểu diễn dưới dạng cây. Trong đó, mỗi nút trong (internal node) biểu diễn một thuộc tính, mỗi nhánh (branch) biểu diễn giá trị có thể có của thuộc tính, mỗi lá (leaf node) biểu diễn các lớp quyết định và đỉnh trên cùng của cây gọi là nút gốc (root).

Cây quyết định được tạo thành bằng cách lần lượt chia (đệ quy) một tập dữ liệu thành các tập dữ liệu con, mỗi tập con được tạo thành chủ yếu từ các phần tử của cùng một lớp. Việc lựa chọn thuộc tính để tạo nhánh của cây được thực hiện thông qua Entropy và Gain.

1.1.1.2 Phân loại cây quyết định

1.1.1.2.1 Cây hồi quy (Regression tree)

Cây hồi quy là cây mà biến phụ thuộc y được ước lượng bởi các hàm có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc điểm trung bình của một học sinh).

1.1.1.2.2 Cây phân loại (Classification tree)

Cây phân loại là cây mà biến phụ thuộc y là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua) hay đánh giá tiến trình học tập của một sinh viên (đúng tiến độ hay không; có thể bị cảnh báo học tập với mức nào). Cây phân loại này cũng chính là cây quyết định sẽ được sử dụng trong bài toán dự đoán tiến trình học tập của sinh viên Đại học Thủy lợi trong luận văn này.

1.1.2 Cấu trúc của cây quyết định

Cây quyết định là một cấu trúc được sử dụng để chia liên tiếp một tập các bản ghi lớn thành các tập con nhỏ hơn bằng cách áp dụng một chuỗi các luật đơn giản. Với mỗi phép chia liên tiếp, các tập con thu được trong tập kết quả sẽ ngày càng giống nhau. Nó có cấu trúc như sau: Mỗi nút mang một thuộc tính (biến độc lập); Mỗi nhánh tương ứng với một giá trị của thuộc tính; Mỗi nút lá là một lớp (biến phụ thuộc)

1.1.3 Xây dựng cây quyết định

1.1.3.1 Phương pháp xây dựng cây quyết định

Việc xây dựng cây quyết định bao gồm 2 giai đoạn: Tạo cây và tỉa cây.

1.1.3.2 Chọn thuộc tính phân tách

Ngay từ khi khởi đầu, tập huấn luyện đã chứa tập các bản ghi mà được phân loại trước - tức là giá trị của biến đích được xác định trong tất cả các trường hợp. Cây quyết định được xây dựng bằng cách phân tách các bản ghi tại mỗi nút dựa trên một thuộc tính đầu vào. Như vậy, rõ ràng nhiệm vụ đầu tiên là phải chọn ra xem thuộc tính nào đưa ra được sự phân tách tốt nhất tại nút đó.

1.1.3.3 Phép kiểm tra để chọn phân tách tốt nhất

Để kiểm tra được thuộc tính nào phân tách tốt nhất sử dụng các độ đo sự đồng nhất như Entropy, Information Gain, Infomation Gain Ratio

1.1.3.3.1 Entropy

Công thức Entropy tổng quát cho một tập mẫu S có C giá trị phân loại:

$$Entropy(S) = \sum_{i=1}^C -p_i \log_2 p_i \quad (2.2)$$

Giả sử phân chia các bộ trong S trên một thuộc tính A bất kỳ, để không mất tính tổng quát có thể xem như A có các giá trị phân biệt $\{a_1, a_2, \dots, a_v\}$, trong đó v là số giá trị phân biệt trong thuộc tính A. Độ đo thông tin có được sau khi phân lớp theo V tập con trên sẽ được tính như sau:

$$Entropy_A(S) = \sum_{j=1}^v \frac{|S_j|}{|S|} \times Entropy(S_j) \quad (2.3)$$

1.1.3.3.2 Infomation gain

Giá trị Gain của thuộc tính A trong tập S được tính như công thức (2.4) dưới đây

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times Entropy(S_v) \quad (2.4)$$

1.1.3.3.3 Information Gain Ratio

Đây là độ đo được mở rộng từ độ đo Information Gain, nó quan tâm đến số lượng và độ lớn của các nhánh khi lựa chọn thuộc tính phân lớp. Thuật toán C4.5 sử dụng độ đo này

để đánh giá sự thay đổi của các thuộc tính. Đại lượng Information Gain Ratio được biểu diễn dưới công thức (2.5) dưới đây

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)} \quad (2.5)$$

SplitInformation(S,A) là thông tin tiềm ẩn được tạo ra bằng cách chia tập dữ liệu trong một số tập con nào đó và được tính theo công thức (2.6).

$$SplitInformation(S, A) = - \sum_{j=1}^v \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \quad (2.6)$$

1.1.3.3.4 Gini Index

1.1.4 Biến đổi cây quyết định thành luật

Sau khi đã xây dựng được mô hình cây quyết định thì có thể biểu diễn tri thức dưới dạng luật IF-THEN. Các luật được xây dựng dựa trên các quy tắc sau:

- Mỗi luật tạo ra từ mỗi đường dẫn từ gốc đến lá.
- Mỗi cặp giá trị thuộc tính dọc theo đường dẫn tạo nên phép kết hợp (phép AND - và)
- Các nút lá mang tên của lớp cần phân loại

1.1.5 Một số thuật toán xây dựng cây quyết định.

1.1.5.1 Thuật toán ID3

1.1.5.1.1 Giới thiệu thuật toán ID3

ID3 xây dựng cây quyết định theo cách từ trên xuống. Nó xác định sự phân lớp các đối tượng bằng cách kiểm tra giá trị của các thuộc tính. Với bất kỳ thuộc tính nào, cũng có thể phân vùng tập hợp các mẫu huấn luyện thành những tập con tách rời, mà ở đó mọi mẫu trong một phân vùng (partition) có một giá trị chung cho thuộc tính đó.

Tại mỗi nút của cây, ID3 chọn một thuộc tính để kiểm tra và dùng kết quả của phép kiểm tra này để phân vùng tập hợp các mẫu thành các phần theo kết quả đó, việc này tiếp tục được thực hiện đệ quy cho đến khi mọi thành viên của phân vùng đều nằm trong cùng một lớp; lớp đó trở thành nút lá của cây.

1.1.5.1.2 Giải thuật Cây quyết định ID3

```

Function induce_tree(tập_ví_dụ, tập_thuộc_tính)
  begin
    if mọi ví dụ trong tập_ví_dụ đều nằm trong cùng một lớp then
      return một nút lá được gán nhãn bởi lớp đó
    else if tập_thuộc_tính là rỗng then
      return nút lá được gán nhãn bởi tuyến của tất cả các lớp trong tập_ví_dụ
    else
      begin
        chọn một thuộc tính P, lấy nó làm gốc cho cây hiện tại;
        xóa P ra khỏi tập_thuộc_tính;
        với mỗi giá trị V của P
          begin
            tạo một nhánh của cây gán nhãn V;
            Đặt vào phân_vùngV các ví dụ trong tập_ví_dụ có giá trị V tại thuộc tính P;
            Gọi induce_tree(phân_vùngV, tập_thuộc_tính), gắn kết quả vào nhánh V
          end
        end
      end
    end
  end

```

1.1.5.1.3 Ví dụ minh họa

1.1.5.2 Thuật toán C4.5

1.1.5.2.1 Giới thiệu thuật toán C4.5

Thuật toán ID3 bị giới hạn bởi việc liên quan đến những thuộc tính mang những giá trị rời rạc rõ ràng, còn những thuộc tính liên tục hoặc những thuộc tính kiểu số thì thuật toán ID3 rất khó xử lý. Trong thuật toán C4.5 sẽ mở rộng phạm vi hoạt động của thuật toán cho những thuộc tính có giá trị liên tục (giá trị số) để phù hợp với thực tế; thuật toán C4.5 đưa ra định nghĩa những giá trị rời rạc mới để phân những giá trị liên tục thành những thuộc tính tượng trưng một lần nữa theo các quy tắc sau:

Thuật toán C4.5 sẽ lựa chọn thuộc tính để phân tách theo nguyên tắc: Tỷ lệ tăng thêm thông tin (GainRatio) cao; Có Entropy của thuộc tính lớn hơn Entropy trung bình của tất cả các thuộc tính.

Một sự cải tiến nữa của của thuật toán C4.5 đó là thuộc tính thiếu giá trị, đây là một vấn đề cũng hay xảy ra trong thực tế. Một cách đơn giản là bỏ đi các mẫu này tuy nhiên nếu có quá nhiều giá trị thiếu hay vai trò của chúng là quan trọng thì sẽ không khả thi.

1.1.5.2.2 Giải thuật Cây quyết định C4.5

Thuật toán tạo cây (S, C)

Bước 1. Tính toán tần suất các giá trị trong các lớp của S

Bước 2. Kiểm tra các mẫu, nếu thuộc cùng một lớp hoặc có rất ít mẫu khác lớp> thì <trả về một nút lá>, ngược lại <Tạo một nút quyết định N>;

Bước 3. Tính giá trị Gain cho mỗi giá trị thuộc tính A

Bước 4. Tại nút N thực hiện kiểm tra để chọn ra thuộc tính có giá trị Gain lớn nhất.

Gọi N.test là thuộc tính có Gain lớn nhất

Bước 5. Nếu N.test là thuộc tính liên tục thì <tìm ngưỡng cho phép tách của N.test

Bước 6. Với tập S thành các tập con S' (được tách theo quy tắc: Nếu N.test là thuộc tính liên tục, tách theo ngưỡng ở bước 5; Nếu N.test là thuộc tính phân loại rời rạc tách theo các giá trị của thuộc tính này)

Bước 7. Nếu S' = rỗng thì gán nút con này của N là nút lá, ngược lại thì gán nút con này là nút được trả về bằng cách gọi đệ quy lại với hàm Create_tree(S') với tập S'

Bước 8. Tính toán lại lỗi của nút N

Bước 9. Return N

1.1.5.2.3 Ví dụ minh họa

1.1.5.3 Đánh giá các thuật toán

1.1.5.3.1 Thuật toán ID3

Giải thuật ID3 là một giải thuật học đơn giản nhưng nó chỉ phù hợp với một lớp các bài toán hay vấn đề có thể biểu diễn bằng ký hiệu. Chính vì vậy, giải thuật này thuộc tiếp cận giải quyết vấn đề dựa trên ký hiệu (symbol - based approach).

Tuy nhiên, giải thuật ID3 cũng có những hạn chế như: Chỉ thích hợp với mô hình có lượng dữ liệu ít, rời rạc; Không thích ứng được với những tập dữ liệu tạp (dễ phát sinh lỗi); Không hiệu quả khi xuất hiện những dữ liệu không mong muốn; Cây quyết định khi dựng ra vẫn còn có thể lớn, rườm rà, chưa được tối ưu ở mức tối đa có thể

1.1.5.3.2 Thuật toán C4.5

C4.5 là sự mở rộng của giải thuật ID3 trên một số khía cạnh, cụ thể là:

Trong việc xây dựng cây quyết định, chúng có thể liên hệ với tập huấn luyện mà có những bản ghi với những giá trị thuộc tính không được biết đến bởi việc đánh giá việc thu thập thông tin hoặc là tỉ số thu thập thông tin, cho những thuộc tính bằng việc xem xét chỉ những bản ghi mà ở đó thuộc tính được định nghĩa.

Trong việc xây dựng cây quyết định, giải thuật C4.5 có thể giải quyết tốt đối với trường hợp giá trị của các thuộc tính là giá trị số và liên tục bằng cách phân ngưỡng đối với các thuộc tính này bằng các phép tách nhị phân. C4.5 cũng có thể giải quyết tốt đối với trường hợp thuộc tính có nhiều giá trị mà mỗi giá trị này lại duy nhất bằng cách đưa vào sử dụng hàm GainRatio. Ngoài ra, trong thuật toán còn có bước tính lỗi cho các nút và cắt tỉa các nhánh không phù hợp

2.2. K Láng giềng gần nhất

2.2.1. Tổng quan về K láng giềng gần nhất

Thuật toán K láng giềng gần nhất (KNN - K Nearest Neighbors) được sử dụng rất phổ biến trong lĩnh vực khai phá dữ liệu. KNN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần phân lớp (Query point) và tất cả các đối tượng trong tập dữ liệu huấn luyện (Training Data).

2.2.2. Thuật toán K láng giềng gần nhất

2.2.2.1. Thuật toán

Bước 1: Xác định giá trị tham số K (số láng giềng gần nhất)

Bước 2: Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong dữ liệu huấn luyện (thường sử dụng khoảng cách Euclidean)

Bước 3: Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với đối tượng cần phân lớp

Bước 4: Lấy tất cả các lớp của K láng giềng gần nhất đã xác định

Bước 5: Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho đối tượng cần phân lớp

2.2.2.2. Hàm tính khoảng cách

Đối với một đối tượng mới cần phân lớp, thuật toán KNN gán phân lớp của đối tượng như một hoặc nhiều đối tượng tương tự nó nhất. Nhưng làm thế nào để định nghĩa được độ tương tự?

Phân tích dữ liệu xác định thước đo khoảng cách để đo độ tương tự. Một thước đo khoảng cách hoặc một hàm khoảng cách d mang giá trị thực, sao cho với bất kỳ tọa độ x , y và z nào thì đảm bảo các tính chất.

Tính chất 1: đảm bảo khoảng cách giữa 2 tọa độ là không âm, và khoảng cách = 0 khi các tọa độ là như nhau.

Tính chất 2: chỉ giao hoán. Khoảng cách từ tọa độ x đến y bằng với khoảng cách từ y đến x

Tính chất 3: là bất đẳng thức tam giác: cho thêm 1 tọa độ thứ 3 thì không bao giờ có thể rút ngắn được khoảng cách giữa 2 tọa độ khác

Hàm khoảng cách phổ biến nhất là hàm khoảng cách Euclide, đây là cách tính khoảng cách thông thường trong thực tế

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (2.7)$$

Với $x = x_1, x_2, \dots, x_m$ và $y = y_1, y_2, \dots, y_m$. Đại diện cho m giá trị thuộc tính của 2 đối tượng x và y (bản ghi).

Ví dụ: Để tính khoảng cách giữa 2 bệnh nhân, bệnh nhân A có thuộc tính tuổi là $x_1 = 20$ và thuộc tính tỷ lệ Na/K là $x_2 = 12$, bệnh nhân B có tuổi là $y_1 = 30$; tỷ lệ Na/K là $y_2 = 8$ khi đó khoảng cách giữa 2 bệnh nhân sẽ là

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} = \sqrt{(20 - 30)^2 + (12 - 8)^2} = 10.77$$

Tuy nhiên, có một vấn đề là khi đo khoảng cách có những thuộc tính có giá trị lớn (ví dụ như thu nhập) sẽ ảnh hưởng nhiều hơn so với các thuộc tính có giá trị nhỏ (ví dụ như số năm công tác). Để tránh điều này người phân tích dữ liệu cần chuẩn hóa (normalize) các giá trị thuộc tính

Đối với các giá trị liên tục có thể sử dụng chuẩn hóa Min – Max hoặc chuẩn hóa Z-Score

Chuẩn hóa Min – Max:

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (2.8)$$

Trong đó:

X : là giá trị thuộc tính của đối tượng

$\min(X)$: là giá trị nhỏ nhất trong miền giá trị của thuộc tính

$\max(X)$: là giá trị lớn nhất trong miền giá trị của thuộc tính

Chuẩn hóa Z-Score:

$$X^* = \frac{X - \text{mean}(X)}{SD(X)} \quad (2.9)$$

Trong đó:

X : là giá trị thuộc tính của đối tượng

$\text{mean}(X)$: là giá trị trung bình trong miền giá trị của thuộc tính;

$SD(X)$: là độ lệch chuẩn của thuộc tính

Đối với các biến là danh sách thì đo khoảng cách bằng Euclide là không thích hợp, ta có thể định nghĩa một hàm khác, sử dụng để so sánh giá trị thuộc tính thứ i của 2 bản ghi như sau:

$$\text{Different}(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{otherwise} \end{cases} \quad (2.10)$$

Trong đó: x_i, y_i là thuộc tính thứ i loại danh sách của đối tượng x và y

Sau đó có thể thay thế hàm $\text{Different}(x_i, y_i)$ cho thuộc tính thứ i trong thước đo khoảng cách Euclide nói trên.

2.2.2.3. Hàm Kết hợp

Sau khi đã có phương pháp xác định độ tương tự của các đối tượng, thì đối với đối tượng chưa được phân loại phải thiết lập như thế nào với các đối tượng tương tự nó để quyết định phân loại cho đối tượng đó, ta cần có một hàm kết hợp [7].

2.2.2.4. Lượng hóa các thuộc tính và kéo dẫn các trục

2.2.2.5. Đánh giá, nhận xét thuật toán

Ngoài việc sử dụng thuật toán KNN để phân loại, nó có thể được sử dụng để đánh giá và dự đoán các biến có giá trị liên tục Một phương pháp để thực hiện điều này là trung bình trọng số cục bộ (Locally weighted averaging).

K láng giềng là thuật toán quan trọng để tiếp cận một CSDL phong phú có nhiều giá trị thuộc tính có sự kết hợp. KNN không chỉ dự đoán các phân loại phổ biến mà còn phân loại được các phân loại hiếm gặp khi có đầy đủ dữ liệu. Vì vậy, tập dữ liệu cần phải được cân bằng, với một tỷ lệ đủ lớn của các phân loại ít phổ biến hơn.

Chọn K

Trong thực tế, không có cách chọn nào được cho là giải pháp tốt nhất. Nếu chọn K nhỏ thì khi phân loại hay dự đoán dễ bị ảnh hưởng bởi các giá trị ngoại lai (nhiều). Còn khi chọn

K quá lớn thì các đặc tính riêng biệt (các láng giềng giống nó nhất) từ tập huấn luyện sẽ bị bỏ qua (làm mịn dữ liệu). Các nhà phân tích dữ liệu cần phải cân đối những cân nhắc khi lựa chọn các giá trị của K .

2.3. Kết luận chương 2

CHƯƠNG III. DỰ BÁO TIẾN TRÌNH HỌC TẬP CỦA SINH VIÊN ĐẠI HỌC THỦY LỢI

1.2 Giới thiệu bài toán

Tại Đại học Thủy lợi, có trường hợp sinh viên học đạt hầu hết các môn (Điểm D tức là từ 4-5.4 điểm theo thang 10 là đạt [1]) nhưng lại không đủ điều kiện làm Đồ án tốt nghiệp (điều kiện làm Đồ án tốt nghiệp là không nợ môn và điểm trung bình chung các môn là 2.0 theo thang điểm 4 [3]), từ đó dẫn đến việc sinh viên bị chậm tiến độ học tập... Hoặc cũng có thể do thiếu hiểu biết mà sinh viên không ý thức được mình sẽ thuộc đối tượng bị cảnh báo học tập...

Do đó, yêu cầu bức thiết đặt ra là cần phải cung cấp thông tin cho sinh viên được biết cảnh báo về tương lai tiến trình học tập của mình như thế nào mà chưa cần phải am hiểu về quy chế đào tạo. Đó là khả năng hoàn thành chương trình học như thế nào? Có đảm bảo tiến độ theo khung chung của nhà trường hay không? Có nguy cơ bị cảnh báo học tập hay không?... Từ đó, sinh viên có thể có những quyết định hợp lý, kịp thời trong thời gian học tập.

1.3 Phân tích và xây dựng mô hình bài toán

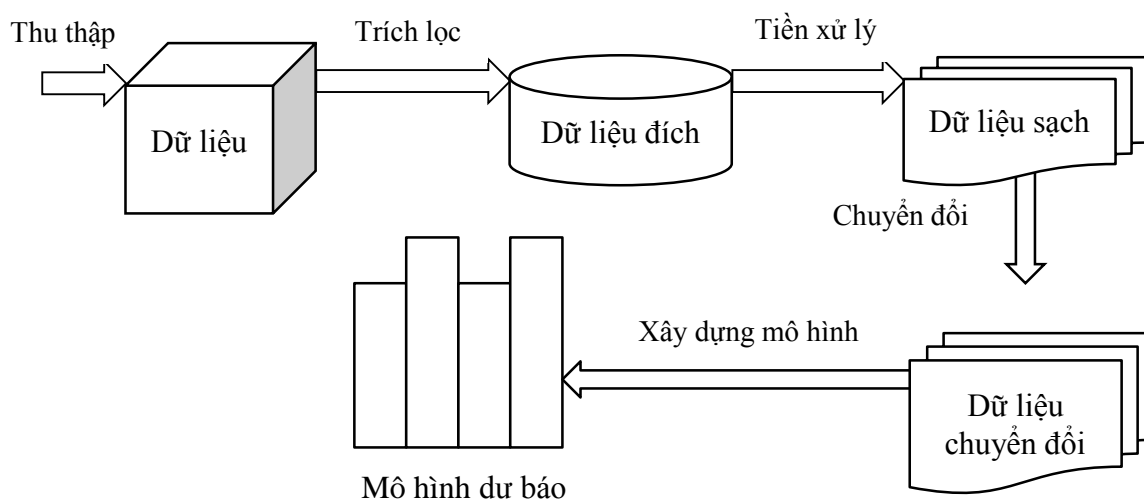
1.3.1 Phân tích bài toán

Trong phạm vi luận văn này, tác giả xác định tiến trình học tập của sinh viên qua 2 câu hỏi xuất phát từ thực tế như sau:

- a. Tiến độ học tập của sinh viên có đúng hạn hay không?
- b. Sinh viên có bị cảnh báo học tập trong quá trình học tập tại trường hay không?

1.3.2 Mô hình bài toán

Mô hình bài toán dự tiến trình học tập của sinh viên Đại học Thủy lợi được thể hiện ở Hình 3.1 dưới đây:



Hình 3.1. Mô hình bài toán

1.3.2.1 Thu thập dữ liệu

Dữ liệu của bài toán được thu thập từ hệ thống phần mềm quản lý đào tạo của Trường Đại học Thủy lợi (phần mềm quản lý và lưu trữ tất cả các thông tin của sinh viên trong trường). Tất cả các dữ liệu được lấy ra dưới định dạng MS Excel từ cơ sở dữ liệu của hệ thống phần mềm này.

1.3.2.2 Trích lọc dữ liệu

1.3.2.3 Tiền xử lý dữ liệu

1.3.2.3.1 Thông tin đầu vào

1.3.2.3.2 Thông tin đầu ra

1.3.2.4 Chuyển đổi dữ liệu

Sau bước tiền xử lý dữ liệu, dữ liệu sạch đang được tác giả chuyển đổi về định dạng CSV, từ đó làm cơ sở dữ liệu chính cho việc thực nghiệm xây dựng mô hình dự báo.

1.3.2.5 Xây dựng mô hình

Tác giả sử dụng phần mềm Weka (Waikato Environment for Knowledge Analysis) để xây dựng mô hình phân lớp và dự đoán.

Tập dữ liệu thực nghiệm gồm 14 tập dữ liệu, mỗi tập dữ liệu có hơn 6000 bản ghi với các thuộc tính là các biến độc lập và biến phụ thuộc gồm: 02 lớp (KHONGDUNG, DUNG) với trường hợp dự báo tiến độ học tập; 05 lớp (KHONG, MOT, HAI, BA, THOIHOC) với trường hợp dự báo mức xử lý học vụ.

Mục tiêu của mô hình là dự báo tiến độ học tập của sinh viên và mức cảnh báo học tập nếu có.

1.4 Thực nghiệm

1.4.1 Phương pháp đánh giá tập dữ liệu

Do tập dữ liệu không nhiều (chỉ có hơn 6000 bản ghi) nên tác giả đã sử dụng phương pháp Cross-validation (hợp lệ chéo – K-fold) để đánh giá.

Tác giả đã thử K với nhiều trường hợp khác nhau và cuối cùng đã sử dụng phân tầng hợp lệ chéo 10-fold để đánh giá độ chính xác phân lớp của bài toán, do kết quả thu được khả quan hơn cả.

1.4.2 Các độ đo được dùng để dự báo

Tác giả đã sử dụng ma trận nhầm lẫn (Confusion Matrix) kết hợp Tỷ lệ đúng tích cực (TP Rate - True Positive Rate; còn gọi là Recall) và Độ chính xác (Precision) để dự báo

Bảng 3.1. Các độ đo dùng để dự báo

Lớp c_i		Được dự báo bởi hệ thống		TP Rate
		Thuộc	Không thuộc	
Dự báo thực sự (đúng)	Thuộc	TP_i	FN_i	$TP_i / (TP_i + FN_i)$
	Không thuộc	FP_i	TN_i	$TN_i / (TN_i + FP_i)$
Precision		$TP_i / (TP_i + FP_i)$	$TN_i / (TN_i + FN_i)$	$(TP_i + TN_i) / (TP_i + TN_i + FP_i + FN_i)$

Trong đó:

TP_i : Số lượng các mẫu thuộc lớp c_i được phân loại chính xác vào lớp c_i

FP_i : Số lượng các mẫu không thuộc lớp c_i bị phân loại nhầm vào lớp c_i

TN_i : Số lượng các mẫu không thuộc lớp c_i được phân loại (chính xác)

FN_i : Số lượng các mẫu thuộc lớp c_i bị phân loại nhầm (vào các lớp khác c_i)

TP Rate (hay Recall): là tỷ lệ đúng tích cực, cho biết tổng số các mẫu thuộc lớp c_i được phân loại chính xác chia cho tổng số các mẫu được phân loại vào lớp c_i .

Precision: là độ chính xác, cho biết tổng số các mẫu thuộc lớp c_i được phân loại chính xác chia cho tổng số các mẫu được phân loại vào lớp c_i

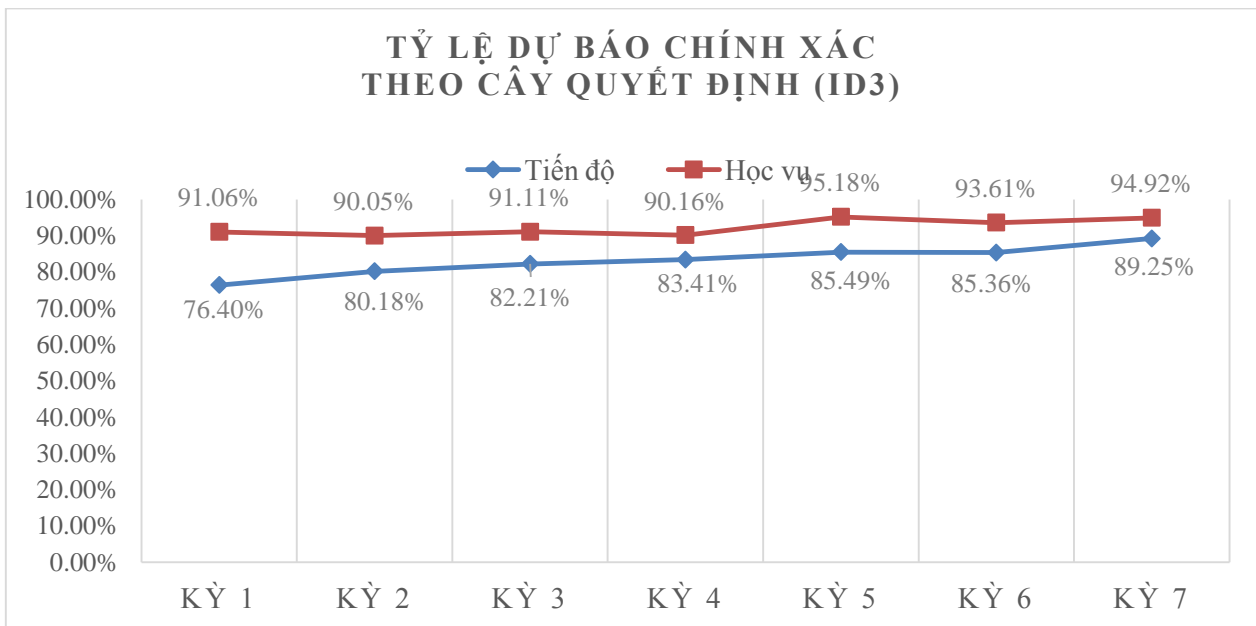
1.4.3 Mô hình dự báo tiến trình học tập của sinh viên

Trong quá trình thực nghiệm dự báo tiến trình học tập của sinh viên, tác giả đã tiến hành thực nghiệm các phương án sử dụng dữ liệu đầu vào khác nhau để đánh giá tính chính xác của mô hình dự báo.

Tác giả nhận thấy rằng: dữ liệu những kỳ trước kỳ n là những dữ liệu gây nhiễu, ảnh hưởng tới kết quả dự báo. Trong bài toàn dự báo của luận văn, dữ liệu gây nhiễu này làm giảm tính chính xác của mô hình dự báo. Tác giả đã tiến hành loại bỏ các dữ liệu gây nhiễu đến quá trình dự báo, và các nội dung tiếp theo của luận văn chỉ đề cập đến kết quả dự báo sau khi đã loại bỏ dữ liệu nói trên.

1.4.3.1 Dự báo bằng Cây quyết định (ID3)

Kết quả dự báo chính xác tiến trình học tập dựa trên dữ liệu của từng học kỳ bằng phương pháp Cây quyết định (ID3) được thể hiện ở hình 3.4 sau:

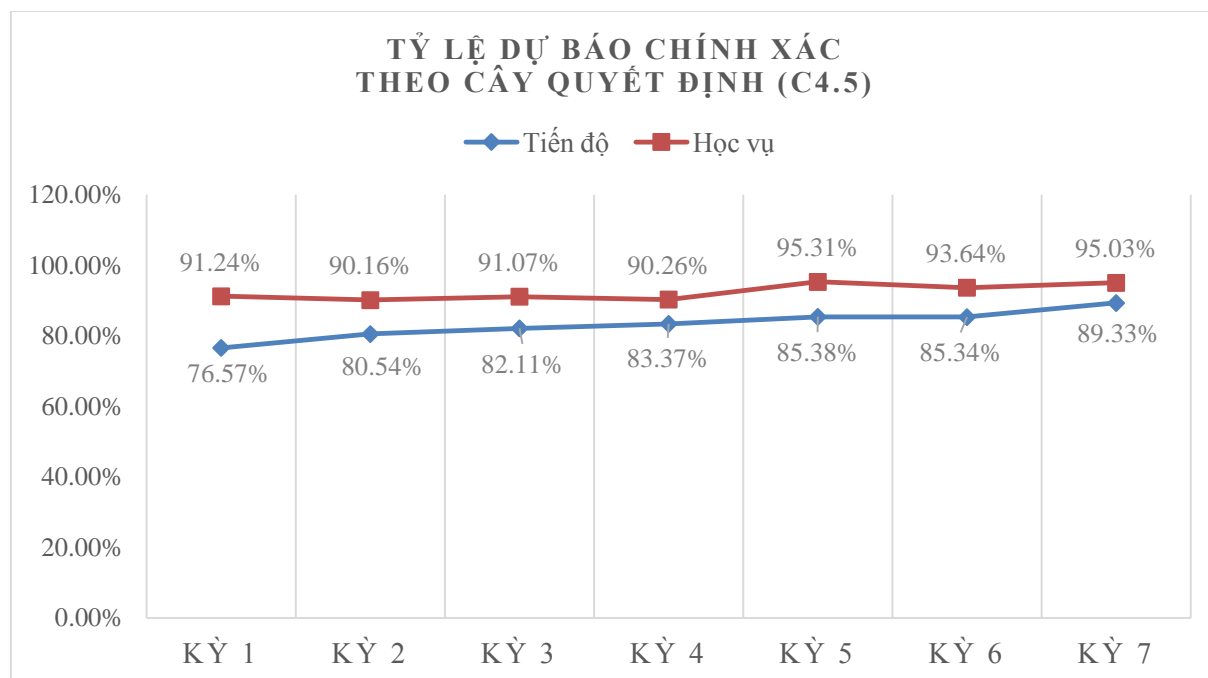


Hình 3.4. Tỷ lệ dự báo chính xác theo Cây quyết định (ID3)

Qua hình 3.4 nói trên, tác giả nhận thấy giải thuật Cây quyết định (ID3) cho kết quả dự báo tốt, đặc biệt là với việc dự báo kết quả xét học vụ của sinh viên (với tỷ lệ dự báo chính xác đều trên 90%). Đối với dự báo tiến độ học tập, tỷ lệ dự báo chính xác sẽ tăng lên khi sinh viên bước vào những học kỳ cuối cùng (từ kỳ 5 trở đi, tỷ lệ dự báo chính xác đều lớn hơn 85%).

1.4.3.2 Dự báo bằng Cây quyết định (C4.5)

Kết quả dự báo chính xác tiến trình học tập (tiền độ và kết quả xét học vụ) dựa trên dữ liệu của từng học kỳ bằng phương pháp Cây quyết định (C4.5) được thể hiện ở hình 3.5 dưới đây.

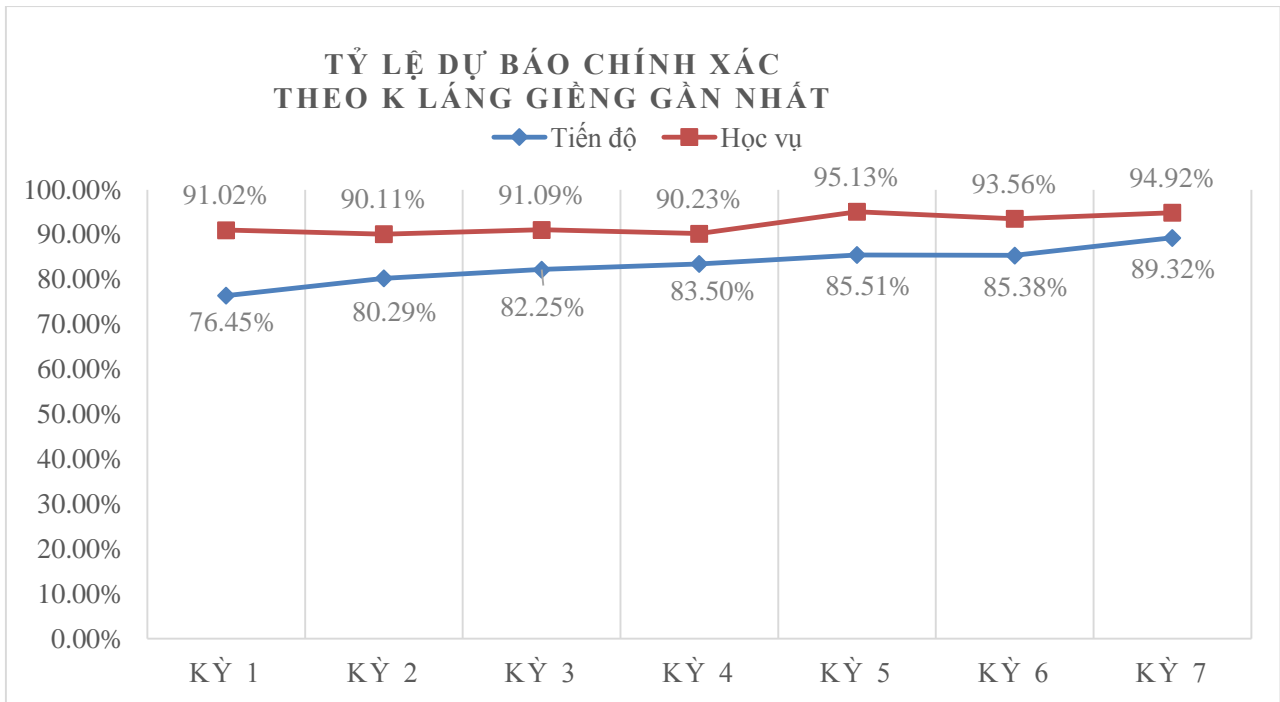


Hình 3.5. Tỷ lệ dự báo chính xác theo Cây quyết định (C4.5)

Qua hình trên tác giả nhận thấy giải thuật Cây quyết định (C4.5) cho kết quả dự báo tốt, đặc biệt là với việc dự báo kết quả xét học vụ của sinh viên (với tỷ lệ dự báo chính xác đều trên 90%). Đối với dự báo tiến độ học tập, tỷ lệ dự báo chính xác sẽ tăng lên khi sinh viên bước vào những học kỳ cuối cùng (từ kỳ 5 trở đi, tỷ lệ dự báo chính xác đều lớn hơn 85%).

1.4.3.3 Dự báo bằng K láng giềng gần nhất (K-NN)

Kết quả dự báo chính xác tiến trình học tập (tiền độ và kết quả xét học vụ) dựa trên dữ liệu của từng học kỳ bằng phương pháp K láng giềng gần nhất được thể hiện ở hình 3.6 như sau:

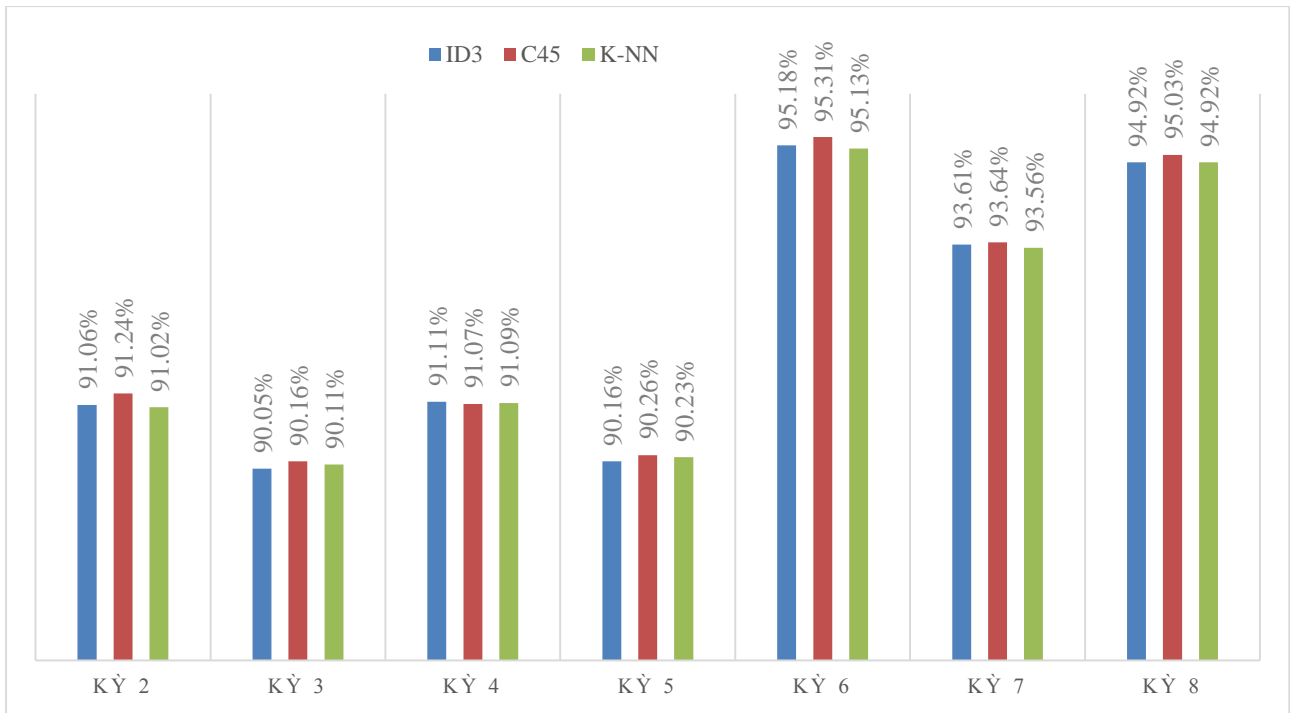


Hình 3.6. Tỷ lệ dự báo chính xác theo K láng giềng gần nhất

Qua hình trên tác giả nhận thấy giải thuật K láng giềng gần nhất cho kết quả dự báo tốt, đặc biệt là với việc dự báo kết quả xét học vụ của sinh viên (với tỷ lệ dự báo chính xác đều trên 90%). Đối với dự báo tiến độ học tập, tỷ lệ dự báo chính xác sẽ tăng lên khi sinh viên bước vào những học kỳ cuối cùng (từ kỳ 5 trở đi, tỷ lệ dự báo chính xác đều lớn hơn 85%).

1.5 Đánh giá thuật toán

Qua thực nghiệm, tác giả nhận thấy kết quả dự báo của cả 3 phương pháp Cây quyết định (ID3, C4.5) hay K láng giềng gần nhất (K-NN) đều có tính chính xác cao cho bài toán dự báo tiến trình học tập (tiến độ học tập và mức cảnh báo học tập) của sinh viên Đại học Thủy lợi. Đặc biệt, sau khi có kết quả học kỳ thứ 5, tính chính xác của dự báo rất cao (trên 85% với dự báo tiến độ và trên 93% với dự báo xử lý học vụ). Có thể nhận biết điều này qua hình 3.7 và hình 3.8.



Hình 3.7. Đánh giá độ chính xác của 3 phương pháp khi dự báo kết quả học vụ



Hình 3.8. Đánh giá độ chính xác của 3 phương pháp khi dự báo tiến độ học tập

1.6 Kết luận chương

KẾT LUẬN

Trong thời đại hiện nay, ứng dụng công nghệ thông tin vào các ngành nghề đang được áp dụng rất rộng rãi, đây thực sự là một công cụ hỗ trợ đắc lực giúp cho con người giải quyết được nhiều vấn đề một cách nhanh chóng, chính xác và hiệu quả trong công việc; một trong những ứng dụng đó là hỗ trợ con người đưa ra các quyết định. Xuất phát từ việc bản thân tác giả là một cán bộ quản lý công tác tổ chức đào tạo tại Trường Đại học Thủy lợi, với mong muốn hỗ trợ tốt nhất cho sinh viên trong việc định hướng học tập của mình trong tương lai; nên tác giả đã lựa chọn nghiên cứu các kỹ thuật khai phá dữ liệu, xây dựng mô hình dự báo để dự báo tiến trình học tập cho sinh viên Đại học Thủy lợi để thực hiện luận văn của mình.

Với tác giả thì khai phá dữ liệu vẫn là một công nghệ mới, việc nghiên cứu trong một thời gian còn ngắn nên vẫn chưa khám phá lĩnh vực hết công nghệ này. Tuy nhiên qua quá trình nghiên cứu luận văn, tác giả đã thu được một số kết quả cũng như nhận thấy một số hạn chế như sau:

1. Kết quả đạt được

Về mặt lý thuyết, tác giả đã có những nghiên cứu về khám phá tri thức và khai phá dữ liệu, các thuật toán dự báo dữ liệu với cây quyết định, K láng giềng gần nhất.

Về mặt thực nghiệm, tác giả đã xây dựng được mô hình dự báo tiến trình học tập của sinh viên Đại học Thủy lợi, từ đó có thể hỗ trợ sinh viên có thể đưa ra các quyết định kịp thời. Tác giả cũng đã có những phân tích, đánh giá được kết quả thực nghiệm

2. Hạn chế

Kết quả mới được thực hiện trên bộ dữ liệu còn chưa đủ lớn (chỉ với dữ liệu hơn 6000 sinh viên khóa đã tốt nghiệp theo học chế tín chỉ tại Trường Đại học Thủy lợi), mô hình dự báo còn đơn giản, mới chỉ sử dụng được mô hình cây quyết định và K láng giềng gần nhất để dự báo. Ngoài ra, do thời gian thực hiện luận văn có hạn nên tác giả chưa xây dựng được một phần mềm để đưa ra kết quả trực quan hơn với đối tượng là sinh viên.

3. Hướng phát triển

Trong thời gian tới, tác giả sẽ tiếp tục nghiên cứu các phương pháp khác nói chung cũng như các thuật toán học máy nói riêng để khai phá dữ liệu với mục tiêu nâng cao độ chính xác việc dự báo. Ngoài ra cũng cần thử nghiệm mô hình với dữ liệu lớn hơn và áp dụng thử cho dữ liệu của một số trường đại học khác.

Ngoài ra, trong thời gian tới tác giả sẽ triển khai xây dựng một ứng dụng để có thể hiển thị kết quả dự báo theo nhiều trường khác nhau để phục vụ tốt hơn cho công tác quản lý của

nhà trường. Mặt khác, cũng sẽ đưa thông tin dự báo tiến trình học tập của từng cá nhân sinh viên Đại học Thủy lợi về tài khoản cá nhân sau khi kết thúc mỗi học kỳ, từ đó sinh viên có thêm thông tin để đưa ra những quyết định phù hợp và kịp thời.

TÀI LIỆU THAM KHẢO

Tài liệu Tiếng Việt

- [1] Bộ Giáo dục và Đào tạo (2007). *Quy chế đào tạo trình độ đại học, cao đẳng theo học chế tín chỉ*, Hà Nội
- [2] Đại học Thủy lợi (2015). *Hướng dẫn thực hiện Quy chế đào tạo đại học và cao đẳng, liên thông cao đẳng lên đại học hệ chính quy theo hệ thống tín chỉ tại Trường Đại học Thủy lợi*, Hà Nội.
- [3] Đại học Thủy lợi (2013). *Quy định về Học phần tốt nghiệp trình độ Đại học, Cao đẳng & Liên thông hệ chính quy và hệ vừa làm vừa học theo hệ thống tín chỉ tại Trường Đại học Thủy lợi*, Hà Nội
- [4] Trần Văn Hiếu (2015). *Khai phá dữ liệu và ứng dụng dự đoán kết quả thi đại học*. Luận văn thạc sĩ công nghệ thông tin. Trường Đại học Sư phạm Hà Nội.
- [5] Khoa Công nghệ thông tin, Trường Đại học Hàng Hải Việt nam (2011). *Bài giảng về Khai phá dữ liệu*, Hải Phòng.
- [6] Khoa Công nghệ thông tin, Trường Đại học Thủy lợi (2011), *Học máy*. Nhà xuất bản Khoa học tự nhiên và công nghệ, Hà Nội.

Tài liệu Tiếng Anh

- [7] Daniel T. Larose, Chantal D. Larose (2014), *Discovering Knowledge in Data: An Introduction to Data Mining*, (Second Edition). John Wiley & Sons Publishers, United States.
- [8] J. Han, M. Kamber, and Jian Pei (2011), *Data Mining Concepts and Techniques*, (Third Edition). Morgan Kaufmann Publishers, United States.