|  |
| --- |
| ĐẠI HỌC QUỐC GIA HÀ NỘI  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**  **SỬ DỤNG MÔ HÌNH PHÂN LỚP ĐỂ DỰ ĐOÁN MẬT ĐỘ GIAO THÔNG**    **Tác giả: Nguyễn Đức Thắng**  **LUẬN VĂN THẠC SĨ**  **Chuyên ngành: HỆ THỐNG THÔNG TIN**  **Hà Nội, 10/2016**  ĐẠI HỌC QUỐC GIA HÀ NỘI  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**  **SỬ DỤNG MÔ HÌNH PHÂN LỚP ĐỂ DỰ ĐOÁN MẬT ĐỘ GIAO THÔNG**    **Tác giả: Nguyễn Đức Thă gs**  Khoa Công Nghệ Thông Tin  Trường Đại học Công Nghệ  Đại học Quốc Gia Hà Nội  **Giảng viên hướng dẫn:**  PGS.TS. Nguyễn Trí Thành  Hà Nội, 10/2016 |

**LỜI CAM ĐOAN**

*“* *Tôi xin cam đoan đây là công trình nghiên cứu của bản thân. Các số liệu, kết quả trình bày trong luận văn này là trung thực và chưa từng được ai công bố trong bất kỳ công trình luận văn nào trước đây.”*

Chữ ký:………………………………………………

**SUPERVISOR’S APPROVAL**

*“I hereby approve that the thesis in its current form is ready for committee examination as a requirement for the Master of Information Systems degree at the University of Engineering and Technology.”*

Chữ ký:………………………………………………

MỤC LỤC

[Danh mục các ký hiệu và chữ viết tắt 4](#_Toc465709397)

[Danh mục bảng 5](#_Toc465709398)

[Danh mục hình vẽ 6](#_Toc465709399)

[MỞ ĐẦU 7](#_Toc465709400)

[NỘI DUNG TRÌNH BÀY 8](#_Toc465709401)

[CHƯƠNG 1. Giới thiệu chung về dự đoán mật độ giao thông 8](#_Toc465709402)

[1.1 Bài toán phân lớp dữ liệu 8](#_Toc465709403)

[1.2 Các bước phân lớp dữ liệu 9](#_Toc465709404)

[CHƯƠNG 2: Tìm hiểu mô hình Decision Tree. 11](#_Toc465709405)

[CHƯƠNG 3: Xây dựng chương trình dựa trên cây quyết định. 13](#_Toc465709406)

[3.1 Mô hình của cây quyết định trong chương trình 13](#_Toc465709407)

[CHƯƠNG 4: Ứng dụng và đưa ra kết quả dự đoán với dữ liệu mẫu 16](#_Toc465709408)

[4.1 Dữ liệu tranning 16](#_Toc465709409)

[4.2 Dữ liệu test 17](#_Toc465709410)

[4.3 Kết quả thực nghiệm 17](#_Toc465709411)

[KẾT LUẬN 19](#_Toc465709412)

[DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN VĂN 20](#_Toc465709413)

[TÀI LIỆU THAM KHẢO 21](#_Toc465709414)

[PHỤ LỤC 22](#_Toc465709415)

# Danh mục các ký hiệu và chữ viết tắt

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Nghĩa đầy đủ** | **Ghi chú** |
| 1 | SVM | Support Vector Machine |  |
|  |  |  |  |
|  |  |  |  |

# Danh mục bảng

[Table 1: So sánh kết quả phân lớp sử dụng SVM, Navies Bayes, J48 và Neural Network. 21](#_Toc460666874)

# Danh mục hình vẽ

[Hình 1:Mô hình phân lớp dữ liệu 10](#_Toc460666833)

[Hình 2: Không gian tuyến tính 12](#_Toc460666834)

[Hình 3: Training Data SVM 13](#_Toc460666835)

[Hình 4: Testing Data SVM 15](#_Toc460666836)

[Hình 5: Kết quả phân lớp sử dụng SVM 15](#_Toc460666837)

[Hình 6: Training Data của Navies Bayes 18](#_Toc460666838)

[Hình 7: Testing Data Navies Bayes 19](#_Toc460666839)

[Hình 8: Kết quả phân lớp sử dụng Navies Bayes 19](#_Toc460666840)

[Hình 9: Training Data của J48 20](#_Toc460666841)

[Hình 10: Testing Data của J48 20](#_Toc460666842)

[Hình 11: Kết quả phân lớp sử dụng J48 21](#_Toc460666843)

[Hình 12: Training Data của Neural Network 21](#_Toc460666844)

[Hình 13: Testing Data của Neural Network 22](#_Toc460666845)

[Hình 14: Kết quả phân lớp sử dụng Neural Network 22](#_Toc460666846)

# NỘI DUNG TRÌNH BÀY

## CHƯƠNG 1. GIỚI THIỆU CHUNG

### Bài toán dự đoán mật độ giao thông

Với tình hình phát triển về kinh tế hiện tại, số lương phương tiện giao thông đặc biệt là ô tô, xe máy đang ngày phát triển với tốc độ lớn về số lượng. Tuy nhiên đi cùng với tốc độ phát triển của phương tiên giao thông thì hạ tầng giao th\ông lại chưa phát triển một cách tương xứng. Tình trạng tắc đường và ùn ứ thường xuyên xảy ra.

Tuy nhiên việc ùn ứ, tắc đường này thường có tính quy luật, ví dụ như tại các thời điểm bắt đầu đi làm buổi sáng hoặc tan tầm là các thời điểm thường xuyên xảy ra tắc đường. Và tại các nút giao thông quan trọng, tình trạng tắc đường cũng thường xuyên xảy ra. Hệ thống có thể dựa vào các đặc điểm, các khoảng thời gian tắc đường, địa điểm hay xảy ra tắc đường để đưa ra các dự báo một cách tương đối chính xác cho những người tham gia giao thông. Những người tham gia giao thông luôn băn khoăn lựa chọn đường đi vào các thời điểm hay xảy ra tắc đường. Nếu chúng ta xây dựng được một hệ thống dự báo tắc đường một cách hiệu quả, người tham gia giao thông có thể dựa vào đó để tìm đường đi thích hợp mà ko bị mất thời gian.

Việc dự đoán mật độ giao thông 3 hướng tiếp cận chính là thông qua quy luật, mô hình hóa và học máy. Trong phương pháp tiếp cận thì học máy có nhiều ưu điểm như không mất thời gian đưa ra các luật, học từ dữ liệu huấn luyện, dễ dàng mở rộng và tái cấu trúc. Các bộ phân loại thường sử dụng trong phương pháp học máy là Support Vector Machine(SVM), Naive Bayes, J48, Neural Network,Maximum Entropy, Decision Tree, Nearest-Neighbors, Sparse Network of Winnows(SNoW).

Trong luận văn nghiên cứu này, tôi sẽ sử dụng và cải tiến mô hình chính là Decision Tree để học dữ liệu mẫu và đưa ra các dự đoán về mật độ giao thông từ các dữ liệu mẫu

### Bài toán phân lớp dữ liệu

Là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp cho trước nhờ một mô hình phân lớp mà mô hình này được xây dựng dựa trên một tập hợp các đối tượng dữ liệu đã được gán nhãn từ trước gọi là tập dữ liệu học (tập huấn luyện).

Quá trình phân lớp còn được gọi là quá trình gán nhãn cho các đối tượng dữ liệu.Như vậy, phân lớp cũng là tiên đoán lại lớp của nhãn.

Có nhiều bài toán phân lớp dữ liệu, như phân lớp nhị phân, phân lớp đa lớp, phân lớp đa trị,…. Phân lớp nhị phân là quá trình tiến hành việc phân lớp dữ liệu vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không một số đặc tính theo quy định của bộ phân lớp. Phân lớp đa lớp là quá trình phân lớp với số lượng lớp lớn hơn hai. Như vậy, tập hợp dữ liệu trong miền xem xét được phân chia thành nhiều lớp chứ không đơn thuần chỉ là hai lớp như trong bài toán phân lớp nhị phân. Về bản chất, bài toán phân lớp nhị phân là trường hợp riêng của bài toán phân lớp đa lớp.

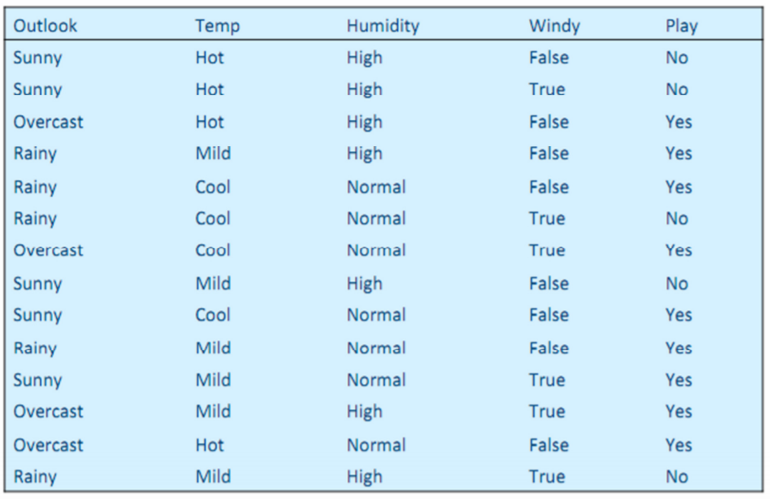
Trong phân lớp đa trị, mỗi đối tượng dữ liệu trong tập huấn luyện cũng như các đối tượng mới sau khi được phân lớp có thể thuộc vào từ hai lớp trở lên.

Với ví dụ là bài toán Dự đoán mật độ Giao thông. Mỗi một đối tượng dữ liệu trong tập huấn luyện là một trường hợp giao thông với một số điều kiện nhất định. Các dữ liệu huấn luyện sẽ không đơn giản chỉ là thông tin tại một thời điểm mà một bộ huấn luyện sẽ ra rất nhiều thời điểm khác nhau. Một lớp chính là một tập hợp các dữ liệu được đánh giá theo giá trị bao gồm từ 1 đến 5 tương ứng với mật độ giao thông từ thưa thớt cho tới đông đúc. Ngoài các bộ huấn luyện cơ bản, còn có thêm các bộ huấn luyện đặc biệt sẽ góp phần tăng độ chính xác khi huyến luyện ví dụ như Tập hợp các ngày lễ tết, Tập hợp các địa điểm đang xây dựng, …. Mỗi một dữ liệu phân lớp sẽ có các giá trị khác nhau, dựa vào các giá trị này , áp dụng các thuật toán phân lớp sẽ phân tích dữ liệu đầu vào và phân các giá trị đó vào các lớp tương ứng.

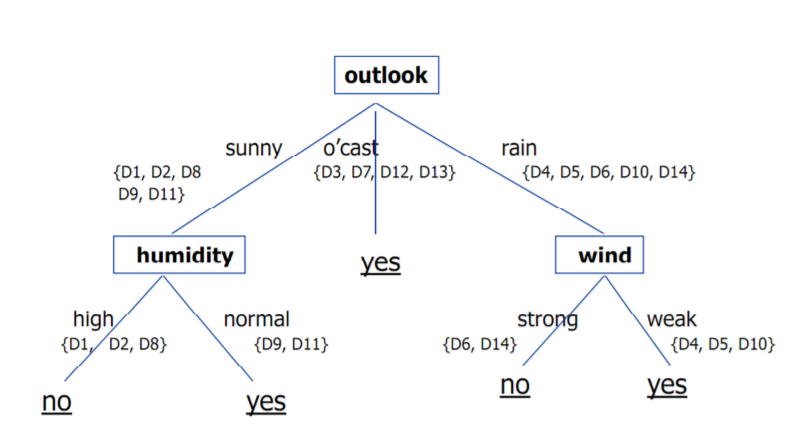
## CHƯƠNG 2: CÁC KHÁI NIỆM TỔNG QUAN.

### 2.1 Mô hình cây quyết định

Cây quyết định (decision tree) là một trong những hình thức mô tả dữ liệu trực quan nhất, dễ hiểu nhất đối với người dùng. Cấu trúc của một cây quyết định bao gồm các nút và các nhánh. Nút dưới cùng được gọi là nút lá, trong mô hình phân lớp dữ liệu chính là các giá trị của các nhãn lớp (gọi tắt là nhãn). Các nút khác nút lá được gọi là các nút con, đây còn là các thuộc tính của tập dữ liệu, hiển nhiên các thuộc tính này phải khác thuộc tính phân lớp. Mỗi một nhánh của cây xuất phát từ một nút p nào đó ứng với một phép so sánh dựa trên miền giá trị của nút đó. Nút đầu tiên được gọi là nút gốc của cây. Xem xét một ví dụ về một cây quyết định như sau[1]:



Từ bảng dữ liệu trên, ta xây dựng được cây quyết định như sau:



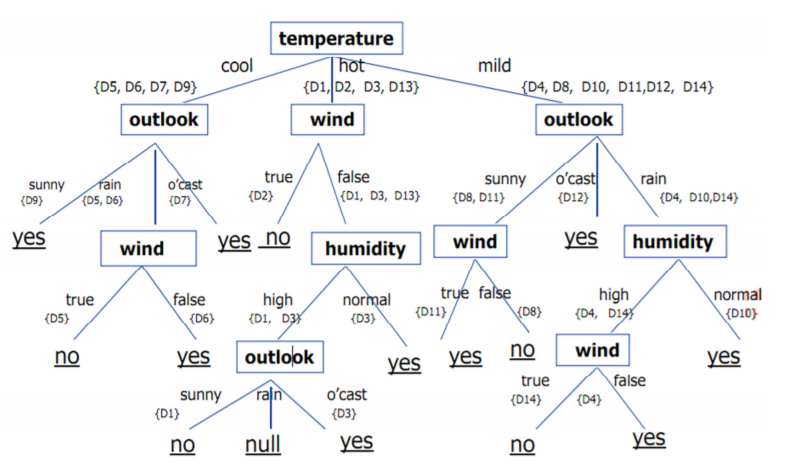
Cây quyết định của ví dụ trên có thể được giải thích như sau: các nút lá chứa các giá trị của thuộc tính phân lớp (thuộc tính “Play”). Các nút con tương ứng với các thuộc tính khác thuộc tính phân lớp; nút gốc cũng được xem như một nút con đặc biệt, ở đây chính là thuộc tính “Outlook”. Các nhánh của cây từ một nút bất kỳ tương đương một phép so sánh có thể là so sánh bằng, so sánh khác, lớn hơn nhỏ hơn… nhưng kết quả các phép so sánh này bắt buộc phải thể hiện một giá trị logic (Đúng hoặc Sai) dựa trên một giá trị nào đó của thuộc tính của nút. Lưu ý cây quyết định trên không có sự tham gia của thuộc tính “thu nhập” trong thành phần cây, các thuộc tính như vậy được gọi chung là các thuộc tính dư thừa bởi vì các thuộc tính này không ảnh hưởng đến quá trình xây dựng mô hình của cây.

Các thuộc tính tham gia vào quá trình phân lớp thông thường có các giá trị liên tục hay còn gọi là kiểu số (ordered or numeric values) hoặc kiểu rời rạc hay còn gọi là kiểu dữ liệu phân loại (unordered or category values). Ví dụ kiểu dữ liệu lương biểu diễn bằng số thực là kiểu dữ liệu liên tục, kiểu dữ liệu giới tính là kiểu dữ liệu rời rạc (có thể rời rạc hóa thuộc tính giới tính một cách dễ dàng).

### 2.2 Chiến lược cơ bản để xây dựng cây quyết định

* Bắt đầu từ nút đơn biểu diễn tất cả các mẫu
* Nếu các mẫu thuộc về cùng một lớp, nút trở thành nút lá và được gán nhãn bằng lớp đó
* Ngược lại, dùng độ đo thuộc tính để chọn thuộc tính sẽ phân tách tốt nhất các mẫu vào các lớp
* Một nhánh được tạo cho từng giá trị của thuộc tính được chọn và các mẫu đƣợc phân hoạch theo.
* Dùng đệ quy cùng một quá trình để tạo cây quyết định.
* Tiến trình kết thúc chỉ khi bất kỳ điều kiện nào sau đây là đúng
* Tất cả các mẫu cho một nút cho trước đều thuộc về cùng một lớp.
* Không còn thuộc tính nào mà mẫu có thể dựa vào để phân hoạch xa hơn.
* Không còn mẫu nào cho nhánh test\_attribute = ai

Tuy nhiên, nếu không chọn được thuộc tính phân lớp hợp lý tại mỗi nút, ta sẽ tạo ca cây rất phức tạp, ví dụ như cây dưới đây:



Như vậy, vấn đề đặt ra là phải chọn được thuộc tính phân lớp tốt nhất. Phần tiếp theo sẽ giới thiệu các tiêu chuẩn, dựa vào các tiêu chuẩn này, ta sẽ chọn ra thuộc tính phân lớp tốt nhất tại mỗi nút.

### 2.3 Thuận lợi và hạn chế của mô hình cây quyết định

**Một số thuận lợi sau đây của cây quyết định được xem như là một công cụ phân loại mà đã chỉ ra trong tài liệu này:**

1. Cây quyết định tự giải thích và khi được gắn kết lại, chúng có thể dễ dàng tự sinh ra. Nói cách khác, nếu cây quyết định mà có số lượng nút lá vừa phải thì người không chuyên cũng dễ dàng hiểu được nó. Hơn nữa, cây quyết định cũng có thể chuyển sang tập luật. Vì vậy, cây quyết định được xem như là dễ hiểu.
2. Cây quyết định có thể xử lý cả thuộc tính tên và số đầu vào.
3. Thể hiện của cây quyết định là đủ đa dạng để biểu diễn cho bất kỳ giá trị rời rạc nào.
4. Cây quyết định có khả năng xử lý các bộ dữ liệu mà có thể gây ra lỗi.
5. Cây quyết định có khả năng xử lý các bộ dữ liệu mà có giá trị rỗng.
6. Cây quyết định được xem như là một phương pháp phi tham số. Điều này có nghĩa là cây quyết định không có giả định về sự phân chia bộ nhớ và cấu trúc phân lớp.

**Bên cạnh đó, cây quyết định cũng có những bất lợi sau đây:**

1. Hầu hết các thuật toán (như ID3 hoặc C4.5) bắt buộc các thuộc tính mục tiêu phải là các giá trị rời rạc.
2. Khi cây quyết định sử dụng phương pháp “chia để trị”, chúng có thể thực hiện tốt nếu tồn tại một số thuộc tính liên quan chặt chẽ với nhau, nhưng sẽ khó khan nếu một số tương tác phức tạp xuất hiện. Một trong những nguyên nhân gây ra điều này là những sự phân lớp mà có mô tả rất mạch lạc về việc phân lớp cũng có thể gặp khó khăn trong việc biểu diễn bằng cây quyết định. Một minh họa đơn giản của hiện tượng này là vấn đề tái tạo cây quyết định (Pagallo và Huassler, 1990). Khi mà hầu hết các cây quyết định phân chia không gian thể hiện thành những khu vực loại trừ lẫn nhau để biểu diễn một khái niệm, trong một số trường hợp, cây nên chứa một vài cây con giống nhau trong thứ tự thể hiện của việc phân lớp. Ví dụ, nếu khái niệm sau mà thể hiện theo hàm nhị phân: y = (A1 ∩ A2) ∪ (A3 ∩ A4) thì cây quyết định đơn biến tối tiểu mà biểu diễn hàm này đã được biểu diễn trong phần 9.3. Lưu ý là cây có chứa 2 bản sao của cùng một cây con.
3. Các đặc tính liên quan của cây quyết định dẫn đến những khó khăn khác như là độ nhạy với tập huấn luyện, các thuộc tính không phù hợp, nhiễu. (Quinlan, 1993).

### 2.7 Cây quyết định mở rộng:

#### 2.7.1 Oblivious Decision Trees

Cây quyết định oblivious là cây quyết định mà tất cả các nút tại cùng cấp thì cùng tính năng. Mặc dù có những hạn chế, song cây quyết định oblivious rất hiệu quả trong việc lựa chọn tính năng. [Almuallim và Deitterich (1994)] cũng như [Schlimmer (1993)] đã đề xuất một thủ tục lựa chọn tính năng trước đây bằng cách xây dựng cây quyết định oblivious, trong khi đó [Langley và Sage (1994)] đã đề nghị lựa chọn ngược cũng sử dụng cây quyết định oblivious. [Kohavi và Sommerfield (1998)] đã chỉ ra rằng cây quyết định oblivious có thể chuyển thành một bảng quyết định. Gần đây [Maimon và Last (2000)] đã đề nghị thuật toán mới IFN (Information Fuzzy Network) để xây dựng cây quyết định oblivious.

**Vì sao phải xây dựng thuật toán IFN?**

• Ưu điểm: - Xây dựng IFN tương tự xây dựng cây quyết định.

- IFN là một đồ thị có hướng chứ không phải là cây.

- IFN sử dụng thông tin chung có điều kiện trong quá trình xây dựng cây, trong khi đó cây quyết định sử dụng số liệu Entropy hoặc Gini.

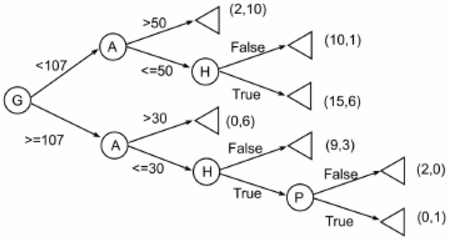
- Chiều cao của IFN không thể vượt quá số lượng đầu vào.

- Các mô hình IFN thường ổn định hơn, điều đó có nghĩa rằng những thay đổi nhỏ trong tập huấn luyện sẽ ảnh hưởng đến nó ít hơn trong các mô hình khác.

• Nhược điểm:

- Tuy nhiên độ chính xác của IFN thấp của cây quyết định.

Ví dụ: Khảo sát một số bệnh nhân có mức đường huyết nhỏ hơn 107 và tuổi lớn hơn 50 thì kết quả nhận thấy rằng: cứ 10 người được chuẩn đoán xem có bị tiểu đường hay không thì 2 người không cần chuẩn đoán bệnh này. Trường hợp khác, khảo sát một số bệnh nhân có đường huyết lớn hơn hoặc bằng 107, tuổi nhỏ hơn hoặc bằng 30, có bị bệnh huyết áp và đang mang thai thì phải làm xét nghiệm tiểu đường. Tương tự cho các đường đi còn lại. Sự khác biệt chính trong cấu trúc của cây quyết định oblivious và cây quyết định thông thường là thứ tự hằng số của thuộc tính đầu vào tại mỗi nút cuối cùng của cây quyết định oblivious. Thuộc tính thứ hai là cần thiết cho việc giảm thiểu toàn bộ tập con của thuộc tính đầu vào (kết quả là giảm kích thước). Các dây cung mà kết nối các nút cuối cùng với các nút của lớp mục tiêu thì được gán nhãn với số lượng mẫu tin phù hợp với đường đi này.

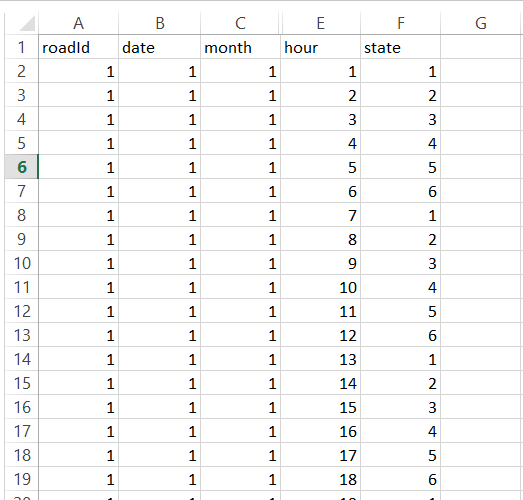


Một cây quyết định oblivious được xây dựng thường xuyên bằng thuật toán tham lam, cái mà cố gắng tối đa hóa các biện pháp thông tin lẫn nhau trong mỗi lớp. Tìm kiếm đệ qui các thuộc tính minh họa, sẽ dừng khi không có thuộc tính nào mà giải thích mục tiêu này với ý nghĩa thống kê.

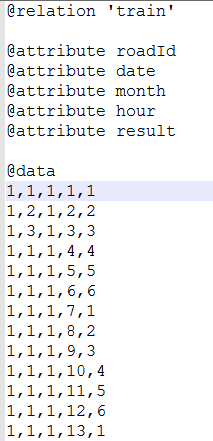
## CHƯƠNG 3: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN MẬT ĐỘ GIAO THÔNG

### 3.1 Mô hình của cây quyết định trong chương trình

Ta có dữ liệu thô có dạng như sau:



Dữ liệu trainning được xây dựng có dạng như sau:



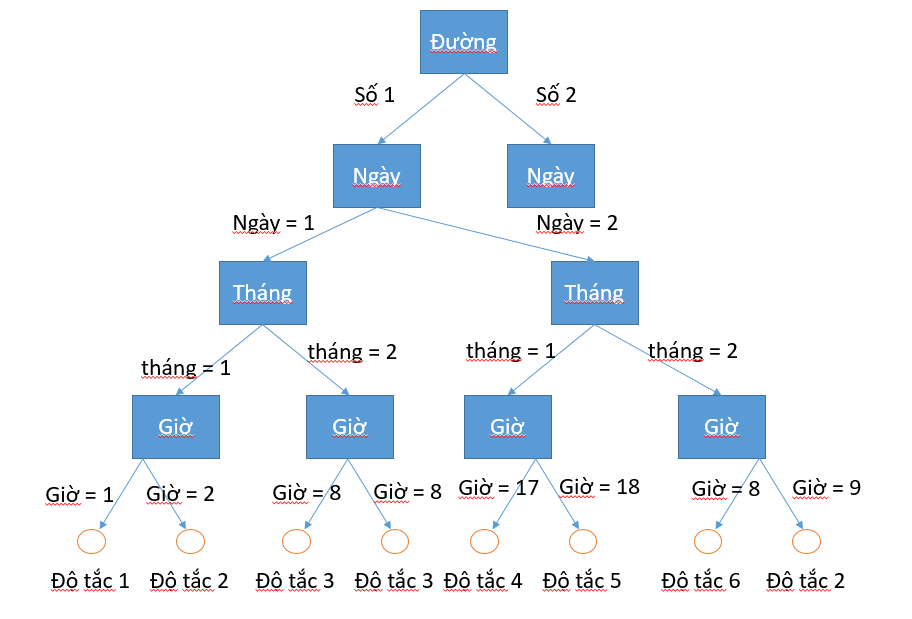
Một bộ dữ liệu tranning bao gồm 3 phần.

Phần 1: Các thông tin mô tả hệ thống (không nhất thiết phải có)

Phần 2: Các Attribute của dữ liệu.

Phần 3: Dữ liệu tranning.

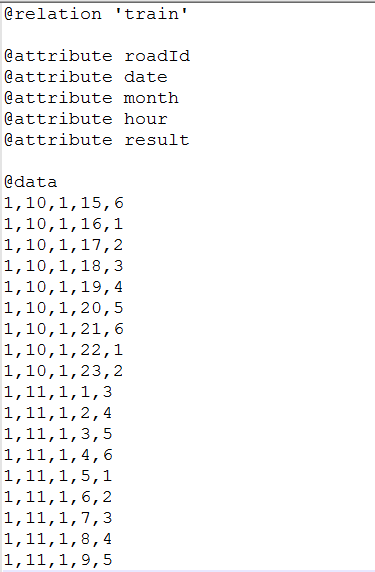
Ta có mô hình cây ứng với bài toán dự đoán mật độ giao thông như sau:



## CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

### 4.1 Dữ liệu tranning

Ta có bộ tranning



Tổng cộng: 8640 dữ liệu traning.

Dữ liệu tranning bao gồm:

RoadId: mã đường

Date: ngày

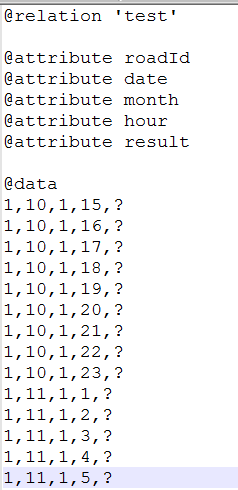
Month: tháng

Hour: Giờ

Result: độ tắc đường.

### 4.2 Dữ liệu test

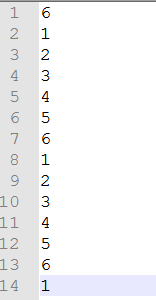
Ta có dữ liệu test như sau



Dữ liệu test có: 1200 test.

### 4.3 Kết quả thực nghiệm

Output:



Đánh giá thực nghiệm:

* Dữ liệu test: 1200
* Trả về kết quả: 1200
* Kết quả chính xác so với tranning: 1200
* Kết quả sai: 0

# KẾT LUẬN

# TÀI LIỆU THAM KHẢO

**Tiếng Việt**

**Tiếng Anh**

1. Naive Bayes Classifiers and Document Classification- Brandon Malone . January 24, 2014
2. Decision Tree Analysis on J48 Algorithm for Data Mining- Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava, Manish Mathuria - Volume 3, Issue 6, June
3. Support Vector Machine (and Statistical Learning Theory) Tutorial Jason Weston  
   NEC Labs America 4 Independence Way, Princeton, USA.  
   [jasonw@nec-labs.com](mailto:jasonw@nec-labs.com)
4. Artifical Neural Networks- Ani1 K. Jain Michigan State University Jianchang M a o K.M. Mohiuddin ZBMAZmadenResearch Center

# PHỤ LỤC

**Phụ lục 4:** **THÔNG TIN LUẬN VĂN THẠC SĨ**

|  |  |
| --- | --- |
| ÐẠI HỌC QUỐC GIA HÀ NỘI  **TRUỜNG ÐẠI HỌC CÔNG NGHỆ** | **CỘNG HÒA XÃ HỘI CHỦ NGHIA VIỆT NAM**  **Ðộc lập - Tự do - Hạnh phúc** |

**THÔNG TIN VỀ LUẬN VĂN THẠC SĨ**

1. Họ và tên học viên: Nguyễn Đức Thắng 2. Giới tính: Nam

3. Ngày sinh:26/03/1991 4. Nơi sinh: Quỳnh Hải, Quỳnh Phụ, Thái Bình.

5. Quyết định công nhận học viên số: ............................, ngày….....tháng.…...nam........

6. Các thay đổi trong quá trình dào tạo: Không

7. Tên đề tài luận van: ......................................................................................................

8. Chuyên ngành: Hệ thống thông tin 9. Mã số: ..................................

10. Cán bộ huớng dẫn khoa học: PGS. TS. Nguyễn Trí Thành

11. Tóm tắt các kết quả của luận van: ..............................................................................

*nêu tóm tắt các kết quả của luận van, nhấn mạnh các* ***kết quả mới*** *nếu có*)

12. Khả năng ứng dụng trong thực tiễn: (*nếu có*) .............................................................

13. Những huớng nghiên cứu tiếp theo: (*nếu có*) .............................................................

14. Các công trình dã công bố có liên quan đến luận van: ...............................................

*liệt kê các công trình theo thứ tự thời gian nếu có)*

|  |  |
| --- | --- |
| *Ngày tháng năm 20*  **Xác nhận của cán bộ huớng dẫn**  *(Kí và ghi rõ họ tên)* | *Ngày tháng năm 20*  **Học viên**  *(Kí và ghi rõ họ tên)* |