BỘ GIÁO DỤC VÀ ĐÀO TẠO
**ĐẠI HỌC ĐÀ NẴNG**

**NGUYỄN VĂN SANG**

# ỨNG DỤNG KHAI THÁC DỮ LIỆU ĐỂ DỰ ĐOÁN SỰ TĂNG TRƯỞNG SỐ THUÊ BAO DI ĐỘNG

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 60.48.01

**TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT**

**Đà Nẵng - Năm 2011**

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. VÕ TRUNG HÙNG**

Phản biện 1: **PGS.TS. PHAN HUY KHÁNH**

Phản biện 2: **GS.TS. NGUYỄN THANH THỦY**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp Thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 11 tháng 09 năm 2011

*Có thể tìm hiểu luận văn tại :*

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

# MỞ ĐẦU

## 1. Lý do chọn đề tài

Ngày nay, viễn thông là một lĩnh vực phát triển rất nhanh chóng. Các công ty viễn thông không ngừng nâng cao, cải tiến chất lượng các dịch vụ của mình, nhằm đáp ứng nhu cầu của khách hàng. Các công ty để tồn tại và phát triển sẽ cạnh tranh với nhau, khách hàng sẽ có nhiều cơ hội để lựa chọn, do đó vấn đề tìm hiểu khách hàng sử dụng các dịch vụ viễn thông rất quan trọng. Trong quá trình như vậy các dịch vụ luôn được thử nghiệm, các hình khuyến mãi đa dạng và phong phú làm cho thị trường viễn thông sôi động ngày càng tăng trưởng nhanh. Tuy nhiên mặt trái của nó cũng làm cho nhà cung cấp dịch vụ rất nhiều phiền phức trong quản lí như thuê bao ảo, tăng trưởng ảo.

Các công ty viễn thông có thể dựa vào CSDL chi tiết cuộc gọi **(**Call Detail Record**)** để phân tích số liệu hành vi sử dụng của khách hàng. Hàng ngày hàng triệu cuộc gọi được ghi nhận tại các tổng đài với mục đích chính là để tính cước cho khách hàng và quản lý mạng, cách mà khách hàng sử dụng mạng, các sản phẩm và các dịch vụ viễn thông. Ngoài ra các công ty viễn thông còn lưu các thông tin khác như phiếu đăng kí dịch vụ, các thông báo lỗi về mạng. Các bản ghi chi tiết cuộc gọi cho biết khi nào thì một dịch vụ được sử dụng mà còn cho biết dịch vụ đó được sử dụng như thế nào.

Một nhà cung cấp dịch vụ thành công khi có quan hệ tốt với khách hàng, giữ được các khách hàng cũ và có thêm khách hàng mới. Thông tin chứa đựng trong các bản ghi cuộc gọi là một tài sản vô cùng quý giá, nó có thể chỉ ra khách hàng cần gì, vì sao mà khách hàng cần các dịch vụ, khách

nào hài lòng, khách hàng nào đem lại lợi nhuận, khách hàng nào có thể rời bỏ. Do đó thách thức lớn nhất là quá trình tìm hiểu hành vi sử dụng của khách hàng để có thể điều chỉnh dịch vụ cũng như đánh giá về sự tăng trưởng giữa ảo và thực.

Người sử dụng không tiếp xúc trực tiếp với nhà cung cấp dịch vụ điện thoại. Khách hàng chỉ tiếp xúc với nhà cung cấp dịch vụ qua bộ phận chăm sóc khách hàng. Do đó nguồn dữ liệu chủ yếu để chúng ta nghiên cứu khách hàng là thông qua các bản tin cuộc gọi.

**2. Mục đích nghiên cứu**

Ứng dụng khai phá dữ liệu để tìm ra những xu hướng của những khách hàng thuê bao, họ có thể rời bỏ mạng viễn thông hay không.

Dự đoán tăng trưởng hằng năm số thuê bao di động để có chính sách điều tiết, đầu tư mạng viễn thông và chăm sóc khách hàng thích hợp.

**3. Phương pháp nghiên cứu**

Phương pháp nghiên cứu tài liệu: Qua nguồn tài liệu được xuất bản, các bài báo đăng trên các tạp chí khoa học, các tài liệu liên quan đến viễn thông.

Phương pháp điều tra: điều tra, thu thập tại các công ty viễn thông.

Phương pháp thực nghiệm: Thực hiện việc cài đặt, thử nghiệm cơ sở dữ liệu, chỉnh sửa để cho kết quả mong đợi.

**4. Ý nghĩa khoa học và thực tiễn**

Kết quả đưa ra có thể đánh giá tình hình thị trường dịch vụ viễn thông hiện nay. Đánh giá được những xu hướng của người sử dụng, ước lượng được bao nhiêu phần trăm thuê bao thực, và thuê bao ảo.

Ước lượng được số thuê bao gia tăng hàng năm.

Kết quả nghiên cứu có thể làm tài liệu cho các nhà cung cấp dịch vụ viễn thông.

**5. Bố cục luận văn**

Luận văn được chia thành 3 chương.

**Chương 1:** Nghiên cứu tổng quan khai phá dữ liệu

Tìm hiểu khái quát chung về khai phá dữ liệu, các bước khai phá dữ liệu, các công cụ cụ thể tiếp cận được đưa ra để giải quyết bài toán.

**Chương 2**: Dự đoán tăng trưởng số thuê bao

Đưa ra bài toán tăng trưởng, phân tích thiết kế hệ thống và đưa ra phương pháp giải bài toán.

**Chương 3**: Xây dựng ứng dụng

Từ kết quả đã nghiên cứu, cài đặt thuật toán xây dựng chương trình ứng dụng có tính thực tiễn.

Đưa ra kết luận, rút ra những mặt ưu điểm và những hạn chế.

# CHƯƠNG 1. NGHIÊN CỨU TỔNG QUAN KHAI PHÁ DỮ LIỆU

## 1.1. KHAI PHÁ DỮ LIỆU

**Định nghĩa**

Khai phá dữ liệu là quá trình tìm kiếm mẫu mới, những thông tin tiềm ẩn mang tính dự đoán dựa vào các khối dữ liệu lớn đã lưu trước đó. Những công cụ KPDL có thể dự đoán những xu hướng trong tương lai, các tri thức mà KPDL mang lại giúp cho các tổ chức ra các quyết địn kịp thời. Sự phân tích một cách tự động và mang tính dự báo của KPDL có ưu thế hơn hẳn so với phân tích thông thường dựa trên những sự kiện mang quá khứ của các hệ hỗ trợ ra quyết định(Decision Support Systems) trước đây.

Với những nội dung được trình bày ở trên, có thể hiểu một cách sơ lược rằng: KPDL được định nghĩa là quá trình tìm kiếm thong tin có ích tiềm ẩn và mang tính dự đoán trong các khối dữ liệu lớn.

**Vai trò của khai phá dữ liệu**

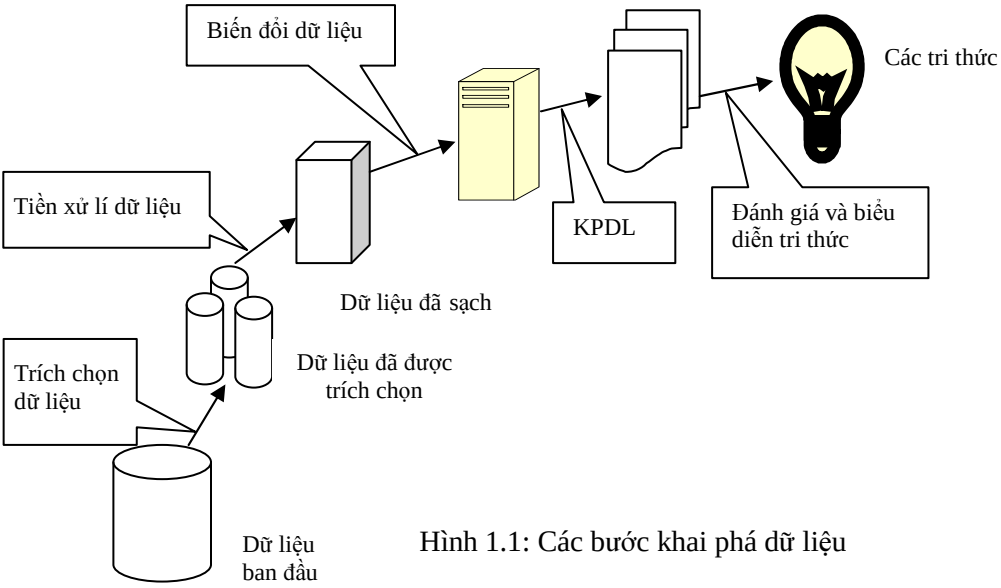
Cuộc cách mạng của khoa học kỹ thuật số cho phép số hóa thông tin trở nên dễ dàng hơn và chi phí lưu trữ từ đó trở nên thấp hơn, số lượng khổng lồ của dữ liệu được tập trung và lưu trữ trong CSDL trên các thiết bị điện tử như: đĩa cứng, băng từ, đĩa quang, CD ROM, thẻ nhớ..khiến tốc độ tăng của dữ liệu quá lớn. Từ đó dẫn đến kỹ thuật thống kê và các công cụ quản trị dữ liệu dựa trên khối dữ liệu khổng lồ đó, không còn phù hợp và không thể phân tích tích đầy đủ nữa.

Dữ liệu của chúng ta sau khi xử lý trực tuyến phục vụ cho một mục đích nào đó được lưu lại ngày càng lớn. Trong khối lượng dữ liệu này còn rất nhiều thông tin có ích mang tính thống kê, có tính quy luật vẫn đang còn tiềm ẩn mà chúng ta chưa biết, đòi hỏi chúng ta cần phải khai phá mới có được. Do đó cần phải có những công cụ tự động rút trích các thông tin, các luật có ích. Một hướng tiếp cận nới có khả năng giúp cho các tổ chức khai thác các thông tin ý nghĩa từ các tập dữ liệu lớn đó là KPDL.

Với những ưu điểm trên, KPDL đã chứng tỏ được tính hữu dụng của nó trong môi trường ngày nay. Vì vậy mà KPDL được ứng dụng rộng rãi trong các lĩnh vực thương mại, tài chính, y học, giáo dục, viễn thông, ngân hàng...

## 1.2. CÁC BƯỚC KHAI PHÁ DỮ LIỆU

KPDL được chia thành các bước như sau:




Hình 1.1: Các bước khai phá dữ liệu

Trích chọn dữ liệu (Data selection): là bước chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (Databases, Data Warehouse).

Tiền xứ lý dữ liệu (Data Preprocessing): là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán..), rút gọn dữ liệu(sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu..), rời rạc hóa dữ liệu (dựa vào histograms, entropy, phân khoảng..) sau bước này, dữ liệu sẽ nhất quán đầy đủ, được rút gọn và được rời rạc hóa.

Biến đổi dữ liệu (Data Transformation): là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai phá ở các bước tiếp theo.

KPDL (Data Mining): đây là bước quan trọng và tiêu tốn nhiều thời gian nhất của KPDL. Áp dụng các kỹ thuật (phần lớn là các kỹ thuật của Machine Learning) để khai phá trích chọn các mẫu (pattern) thông tin dựa vào các mối liên hệ đặc biệt trong dữ liệu

Đánh giá và biểu diễn tri thức (Knowledge Representtation & Evaluation):

Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin và mối liên hệ đặc biệt trong dữ liệu đã được khai phá, biểu diễn theo dạng gần gũi với người sử dụng như đồ thịm cây, bảng biểu, luật.. đồng thời bước này cũng đánh giá những tri thức khai phá được theo những tiêu chí nhất định.

Trong giai đoạn KPDL, có thể cần sự tương tác của người dùng để điều chỉnh và rút ra các tri thức cần thiết.

## 1.3. CÁC DẠNG DỮ LIỆU ĐƯỢC KHAI PHÁ

KPDL đã chứng tỏ được những tính hữu dụng trong thực tế và vì vậy mà được ứng dụng rộng rãi trong các lĩnh vực thương mại, tài chính, y học, giáo dục, viễn thông, ngân hàng.. với những CSDL đã có để đưa ra những luật. KPDL có khả năng chấp nhận một số dạng CSDL như sau:

CSDL giao tác (Transactonal Databases): là dạng dữ liệu tác nghiệp có các bản ghi giao tác. Dạng CSDL này phổ biến trong ngân hàng.

CSDL quan hệ (Relational Databases): là dạng dữ liệu tác nghiệp được tổ chức theo mô hình dữ liệu quan hệ.

CSDL đa chiều (Mutidimention Structures, Data Warehouses): là các kho dữ liệu được tập hợp và chọn lọc từ nhiều nguồn dữ liệu khác nhau. Dạng dữ liệu này chủ yếu phục vụ cho quá trình phân tích cung như khai phá tri thức và hỗ trợ quá trình ra quyết định

CSDL quan hệ-hướng đối tượng (Object Relational Databases): là dạng dữ liệu lai giữa hai mô hình quan hệ và đối tượng.

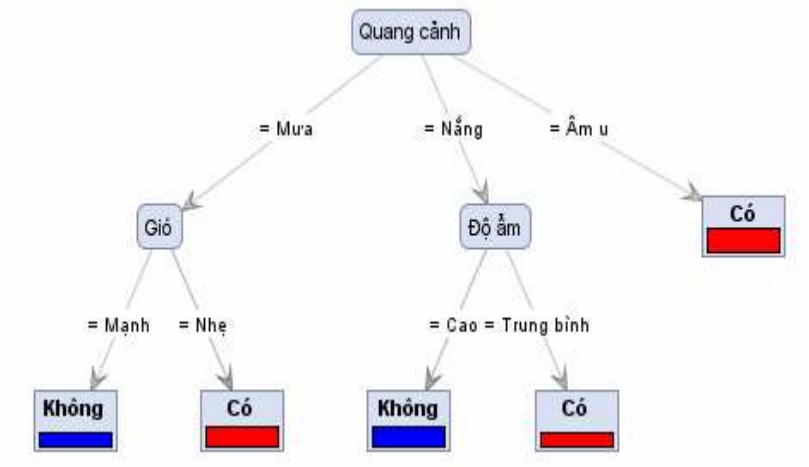
## 1.4. HƯỚNG TIẾP CẬN VÀ KỸ THUẬT KHAI PHÁ DỮ LIỆU

KPDL là một lĩnh vực rộng với nhiều hướng nghiên cứu, tiếp cận khác nhau. Một số hướng tiếp cận chính của KPDL được phân chia theo chức năng theo lớp các bài toán khác nhau.

### 1.4.1. Cây quyết định và luật

Cây quyết định là một phương pháp mô tả tri thức dạng đơn giản nhằm phân các đối tượng dữ liệu thành một số lớp nhất định. Các nút của cây được gán nhãn là tên các thuộc tính, các cạnh được gán các giá trị của các thuộc tính, các lá miêu tả các lớp khác nhau. Các đối tượng được phân

lớp theo các đường đi trên cây, qua các cạnh tương ứng với giá trị của thuộc tính các đối tượng lá.[1]




Hình 1.2: Mô tả cây quyết định

Cây quyết định trên miêu tả điều kiện chơi thể thao với các thuộc tính đặt ra quang cảnh, gió, độ ẩm. Có hai giá trị lá "Có" và "Không".

Cây quyết định và luật có ưu điểm là hình thức miêu tả đơn giản, mô hình suy diễn khá dễ đối với người sử dụng. Tuy nhiên, giới hạn của nó là miêu tả cây và luật chỉ có thể biểu diễn được một số dạng chức năng, vì vậy giới hạn cả về độ chính xác và mô hình, Cho đến nay đã có rất nhiều giải thuật suy diễn sử dụng các luật và cây quyết định được áp dụng trong máy học và thống kê.

### 1.4.2. Phân lớp Bayes

Lý thuyết Bayes cung cấp một tiếp cận theo xác xuất để suy diễn. Nó dựa trên giả thuyết rằng số lượng của khuynh hướng bị chi phối  bởi phân

bố xác suất và  quyết định tối ưu có thể được tạo bởi sự suy luận về những xác suất đi liền với dữ liệu được quan sát. Đây là vấn đề quan trọng của máy học bởi vì nó cung cấp một tiếp cận định lượng cho việc xem xét cẩn thận bằng chứng hỗ trợ những giả thuyết thay đổi. Lý thuyết Bayes cung cấp giải thuật học cơ bản mà vận dụng những xác xuất cũng như là một khung làm việc cho sự phân tích sự hoạt động của những giải thuật mà không thể vận dụng rõ ràng .

Học theo xác suất: Tính xác suất hiện cho giả thuyết, trong số những tiếp cận thực dụng nhất cho các kiểu chắc chắn của những vấn đề học.

Tính tăng dần: mỗi ví dụ huấn luyện có thể gia tăng việc tăng hoặc giảm mà không gian giả thuyết đúng. Kiến thức trước có thể kết hợp với dữ liệu được quan sát.

Tiên đoán xác suất: Tiên đoán nhiều không gian giả thuyết, được đo bởi xác suất của nó.

### 1.4.3. Hồi quy

Hồi quy - nói theo cách đơn giản, là đi ngược lại về quá khứ (regression) để nghiên cứu những dữ liệu (data) đã diễn ra theo thời gian (dữ liệu chuỗi thời gian - time series) hoặc diễn ra tại cùng một thời điểm (dữ liệu thời điểm hoặc dữ liệu chéo - cross section) nhằm tìm đến một quy luật về mối quan hệ giữa chúng. Mối quan hệ đó được biểu diễn thành một phương trình (hay mô hình) gọi là: phương trình hồi quy mà dựa vào đó, có thể giải thích bằng các kết quả lượng hoá về bản chất, hỗ trợ củng cố các lý thuyết và dự báo tương lai.

Trong phân tích hoạt động kinh doanh cũng như trong nhiều lĩnh vực khác, hồi quy là công cụ phân tích đầy sức mạnh không thể thay thế, là phương pháp thống kê toán dùng để ước lượng, dự báo những sự kiện xảy ra trong tương lai dựa vào quy luật quá khứ

***1.4.3.1. Phương pháp hồi quy đơn***

Còn gọi là hồi quy đơn biến, dùng xét mối quan hệ tuyến tính giữa 1 biến kết quả và 1 biến giải thích hay là biến nguyên nhân (nếu giữa chúng có mối quan hệ nhân quả). Trong phương trình hồi quy tuyến tính, một biến gọi là: biến phụ thuộc; một biến kia là tác nhân gây ra sự biến đổi, gọi là biến độc lập.

***1.4.3.2. Phương pháp hồi quy bội***

Còn gọi là phương pháp hồi quy đa biến, dùng phân tích mối quan hệ giữa nhiều biến số độc lập (tức biến giải thích hay biến nguyên nhân) ảnh hưởng đến 1 biến phụ thuộc (tức biến phân tích hay biến kết quả).

# CHƯƠNG 2. DỰ ĐOÁN TĂNG TRƯỞNG SỐ THUÊ BAO

## 2.1. GIỚI THIỆU BÀI TOÁN

### 2.1.1. Bài toán

Trong lĩnh vực viễn thông số lượng khách hàng sử dụng dịch vụ thuê bao rất lớn. Đặc biệt trong thời gian gần đây với sự phát triển nhanh của ngành này kèm theo là các chính sách quản lý thông thoáng đã tạo điều kiện cho người dùng thỏa mãn nhu cầu sử dụng. Bên cạnh đó các hình thức khuyến mãi mang tính cạnh tranh lành mạnh được triển khai nhằm thu hút khách hàng về mình. Từ đó nảy sinh mặt trái là thuê bao ảo, một số người dùng nhiều thuê bao chỉ mục đích là tận dụng chính sách khuyến mãi. Để nhìn nhận vấn đề một cách khách quan hơn chúng ta dùng công cụ khai phá dữ liệu để phân tích những khách hàng đâu là tiềm tàng gắn bó, đâu là thuê bao ảo và sẽ rời bỏ, từ đó có thể điều chỉnh chính sách hợp lý và cuối cùng là có thể dự đoán sự tăng trưởng hàng năm của số lượng thuê bao.

### 2.1.2. Các tập CSDL quản lí thuê bao liên quan đến bài toán

#### *2.1.2.1. Giới thiệu về chi tiết cuộc gọi*

#### *2.1.2.2. CSDL tính cước (Billing')*

#### *2.1.2.3.Cơ sở dữ liệu khách hàng (Customer)*

### 2.1.3. Một số thuộc tính của Chi tiết cuộc gọi

## 2.2. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

### 2.2.1. Cách giải quyết yêu cầu của bài toán

### 2.2.2. Phương pháp triển khai

### 2.2.3. Nội dung triển khai

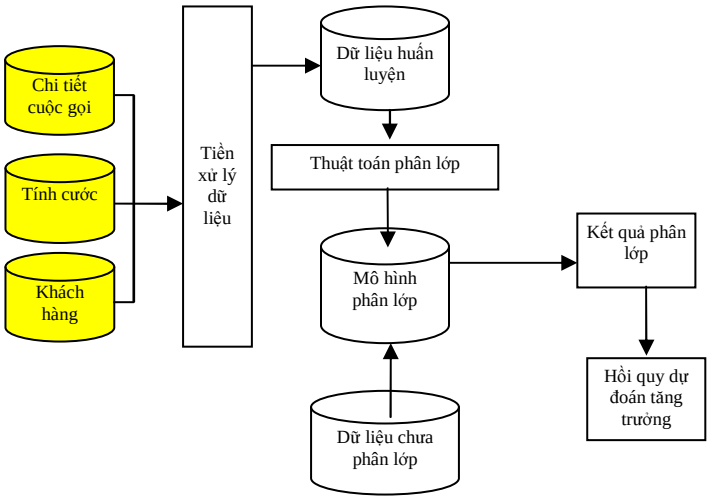
### 2.2.4. Xây dựng tập CSDL huấn luyện

### 2.2.5. Công nghệ sử dụng

### 2.2.6. Các công việc tiến hành với dữ liệu

### 2.2.7. Phân tích thiết kế hệ thống

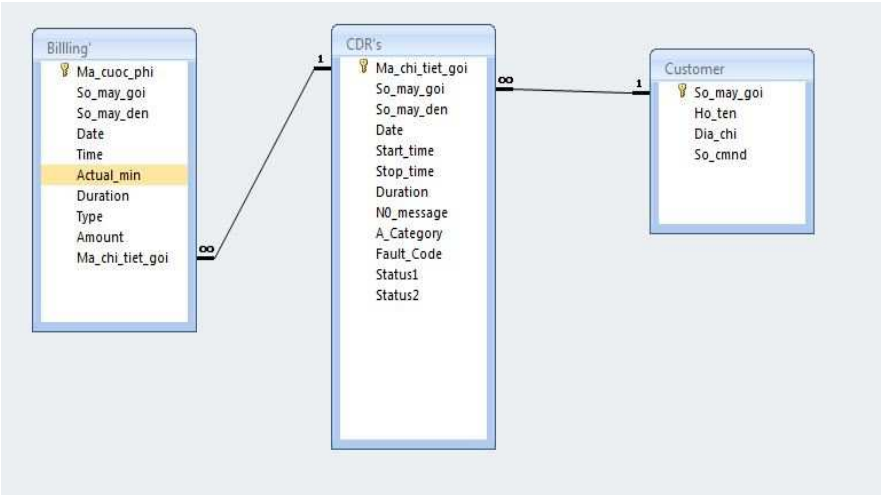
#### *2.2.7.1. Kiến trúc hệ thống*




Hình 2.1: Mô hình hệ thống

#### *2.2.7.2. Các bảng dữ liệu*

#### *2.2.7.3 lược đồ quan hệ của các đối tượng*




Hình 2.2: Mô hình quan hệ giữa các CSDL

## 2.3. PHÂN LỚP DỰ ĐOÁN XU HƯỚNG KHÁCH HÀNG

### 2.3.1. Xây dưng tập dữ liệu huấn luyện (Training Data)

Dựa trên các cơ sở dữ liệu đã có như Chi tiết cuộc gọi, tính cước, thông tin về khách hàng chúng ta chọn các thuộc tính chứa nhiều thông tin có khả năng mang lại cho việc dự đoán để chúng ta tạo ra cơ sở dữ liệu tổng hợp. Cơ sở dữ liệu này được tạo ra từ những thông tin quan trọng nhất và có thể dễ dàng phân tích cho việc dự đoán, được gọi là tập là DL đầu vào hay còn gọi là tập DL huấn luyện (Training data).

Bảng 2.1: Tập CSDL huấn luyện

| Tên khách hàng | Số thuê bao | Sử dụng DV | Thời gian gọi | Tin nhắn | Tài khoản | Rời bỏ |
|---|---|---|---|---|---|---|
| Nguyễn Văn An | 0905245678 | Nhiều | Vừa | Nhiều | Cao | Không |
| Lê Thanh Bình | 0935234532 | Ít | Ngắn | Ít | Thấp | Có |
| Lê Trung Kiên | 01223563456 | Trung bình | Dài | Trung bình | Trung bình | Không |
| Thái Xuân Lan | 0903541789 | Ít | Ngắn | Ít | Cao | Có |
| Đỗ Kim Lan | 0904237865 | Nhiều | Dài | Nhiều | Thấp | Không |
| Trần Thúy Hằng | 0932456654 | Nhiều | Vừa | Trung bình | Trung bình | Không |
| NguyễnVăn Nam | 01215673565 | Trước | Ngắn | Ít | Cao | Có |
| Lê Hải Nam | 0905234561 | Trước | Dài | Nhiều | Trung bình | Không |

### 2.3.2. Giới thiệu về phân lớp

#### *2.3.2.1. Xây dựng mô hình*

#### *2.3.2.2. Sử dụng mô hình*

### 2.3.3. Một số phương pháp phân lớp

### *2.3.3.1. Xây dựng cây quyết định*

Xây dựng cây quyết định là vấn đề then chốt và quan trọng nhất của việc khai phá dữ liệu bằng kỹ thuật này. Các thuật toán xây dựng cây quyết định đã được các nhà khoa học phát triển, công bố và giới thiệu. Một số thuật toán tiêu biểu như sau:[4]

**Xây dựng cây**

**Thuật toán tổng quát xây dựng cây quyết định**

Trong khai phá dữ liệu bằng cây quyết định thì xây dựng cây là vấn đề mấu chốt và quan trọng nhất. Các thuật toán xây dựng cây quyết định đã được các nhà khoa học phát triển, công bố và cải tiến theo thời gian. Tuy nhiên, về mặt tổng quát thì một cây quyết định được xây dựng theo thuật toán sau:

*Dữ liệu vào:* Tập dữ liệu D, tập danh sách thuộc tính, tập nhãn lớp

*Dữ liệu ra:* Mô hình cây quyết định

*Thuật toán:* Tạocây (Tập dữ liệu E, tập danh sách thuộc tính F, tập nhãn lớp)

1 Nếu điều_kiện_dừng (E,F) = đúng

2 nútlá = CreateNode ()

3 nútlá.nhãnlớp=Phânlớp (E)

4 return nútlá

5 Ngược lại

6 Nútgốc = CreateNode ()

7 Nútgốc.điềukiệnkiểmtra = tìm_điểm_chia_tốt_nhất (E, F)

8 Đặt V = {v| v thoả điều kiện là phần phân chia xuất phát từ Nútgốc}

9 Lặp qua từng tập phân chia v ∈ V

10 Đặt $E_v$ = {e | Nútgốc.điềukiệnkiểmtra(e) = v và e ∈ E}

Đặt F = F \ {các giá trị của điều kiện để phân chia v}

11 Nútcon = Tạocây ($E_v$, F, tập nhãn lớp)

12 Dừng lặp

13 End if

14 Trả về nútgốc.

*Hàm chính*

Gọi hàm Tạocây (Tập dữ liệu E, tập danh sách thuộc tính của E, tập nhãn lớp).

#### *2.3.3.2. Phân lớp Bayes*

### 2.3.4. Dự đoán sự tăng trưởng

#### *2.3.4.1. Phương pháp hồi qui tuyến tính*

Phân tích hồi qui tuyến tính là một mô hình dự báo thiết lập mối quan hệ giữa biến phụ thuộc với hai hay nhiều biến độc lập. Trong phần này, chúng ta chỉ xét đến một biến độc lập duy nhất. Nếu số liệu là một chuỗi theo thời gian thì biến độc lập là giai đoạn thời gian và biến phụ thuộc thông thường là doanh số bán ra hay bất kỳ chỉ tiêu nào khác mà ta muốn dự báo. Mô hình này có công thức:Y = ax + b[6]

$$a = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$$

Trong đó : y - Biến phụ thuộc cần dự báo.

x - Biến độc lập

a - Độ dốc của đường xu hướng

b - Tung độ gốc

n - Số lượng quan sát

#### *2.3.4.2. Mô hình dự báo theo phương trình hồi quy*

# CHƯƠNG 3. XÂY DỰNG ỨNG DỤNG

## 3.1. GIỚI THIỆU

Phần mềm được xây dựng với cho phép tương tác với dữ liệu và thực hiện khai phá dữ liệu. Để tương tác với dữ liệu phần mềm thực hiện các chức năng như cập nhật, khởi tạo, và xem dữ liệu gốc. Trong quá trình khai phá dữ liệu thực hiện nhiệm vụ chính của bài toán đó là dự đoán những khách hàng rời bỏ và dự đoán tăng trưởng số thuê bao hằng năm.

Trong bài toán này đưa ra đó là phân lớp khách hàng dựa trên hai xu hướng đó là gắn bó và rời bỏ, từ đó xác định được mối tương quan giữa giá trị "thực" và "ảo" phục vụ riêng cho từng lớp khách hàng có cùng nhu cầu, sở thích, đưa ra các chính sách giá ưu đãi và các chương trình khuyến mãi đối với từng lớp đối tượng riêng. Chẳng hạn, đối với người dùng điện thoại di động trả trước, có người gọi đi nhiều nhưng có người hầu như chỉ sử dụng để nhận các cuộc gọi thì chính sách đối với hai đối tượng này như thế nào? Người gọi nhiều có nhu cầu giá cước thấp, ta có thể đưa ra chương trình giảm giá cước từ phút gọi thứ bao nhiêu trở đi. Nhưng đối với người dùng chỉ nghe thì chương trình này không có ý nghĩa với họ mà chương trình tăng ngày sử dụng sẽ có ý nghĩa hơn.

Sau khi ta có được dự đoán được những giá trị thực của thuê bao ta dùng phương pháp phân tích hồi quy để dự đoán tăng trưởng hằng năm. Các con số và giá trị được đưa ra giúp nhà cung cấp dịch vụ đánh giá khách quan về mặt định lượng số thuê bao. Các số liệu được đưa ra truy xuất dưới dạng biểu đồ và dạng bảng. Những con số được đưa ra minh họa, phản ánh

thực tế sự tăng trưởng của số thuê bao di động. Phần mềm thực hiện hai chức năng dự đoán trên có ý nghía thực sự bởi nhà cung cấp dịch vụ viễn thông, là một cách để đi tìm những chế và hiệu quả sau hằng năm hoạt động để có những phương pháp chính sách điều chỉnh phù hợp hơn.

## 3.2. QUÁ TRÌNH PHÁT TRIỂN

Đề tài dựa trên ý tưởng dựa vào khai phá dữ liệu để phân tích và chăm sóc khách hàng viễn thông. Trong đó dựa vào những thông tin, tính chất của khách hàng lưu trên cơ sở dữ liệu quản lý để tổng hợp thành một cơ sở dữ liệu mới. Phân lớp khách hàng còn là đầu vào cho rất nhiều bài toán khác nữa mà dưới đây là một ví dụ đối với kho dữ liệu cước điện thoại của công ty VMS Mobifone. Đây cũng chính là việc áp dụng thử nghiệm việc phân lớp sử dụng cây quyết định trong khuôn khổ luận văn này. Bài toán đặt ra phân tích những đặc trưng của ngành viễn thông và công cụ khai phá dữ liệu để phân tích xu hướng, dự đoán những người có khả năng rời bỏ và dự đoán tăng trưởng số thuê hằng năm.

Trong quá trình thực hiện đề tài dưới sự tham khảo và tìm hiểu của công ty VMS Mobifone, căn cứ vào các giá trị về cuộc gọi chi tiết và bảng tính tiền và quản lí khách hàng để đưa ra tập dữ liệu huấn luyện. Áp dụng thuật toán phân lớp cây quyết định và Bayes để đưa ra phân tích xu hướng của mỗi khách hàng. Sử dụng phân tích hồi quy để dự đoán sự tăng trưởng số thuê bao hằng năm. Trong khuôn khổ của đề tài này

## 3.3. XÂY DỰNG DEMO

### 3.3.1 Giao diện chính


Hình 3.1: Bảng nhập dữ liệu và kết quả giá trị dự đoán

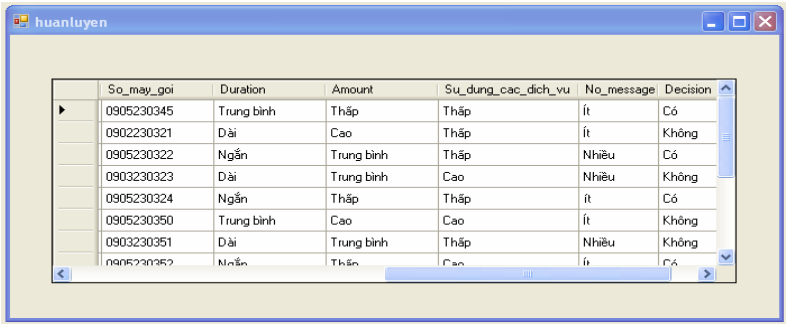
Menu chính của phần mềm dự đoán tăng trưởng phía bên trái thực hiện tương tác với CSDL bao gồm cập nhật, khởi tạo, xem dữ liệu gốc. Bên phải là thao tác với dự đoán. Dự đoán xu hướng là dự đoán đưa ra danh sách những người có thể rời bỏ và không rời bỏ.

Dự đoán tăng trưởng là dự đoán số thuê bao tăng hàng năm là bao nhiêu, có thể tính theo phần trăm được xem dưới dạng bảng và biểu đồ.

### 3.3.2. Menu cập nhật dữ liệu

### 3.3.3. Xem dữ liệu gốc

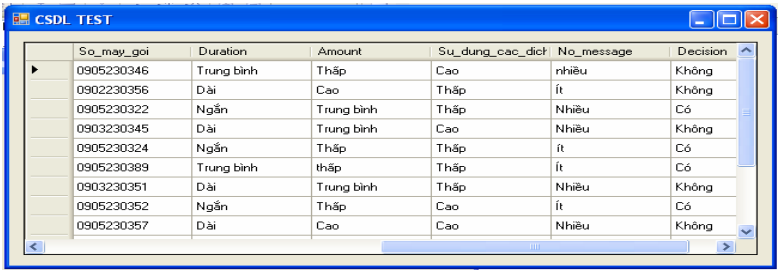
### 3.3.4. Cơ sở dữ liệu huấn luyện




Hình 3.4: Bảng cơ sở dữ liệu huấn luyện

Là tập cơ sở dữ liệu dựa vào những thuộc tính chính mà thuê bao sẽ rời bỏ hay không.

### 3.3.5. Cơ sở dữ liệu Test

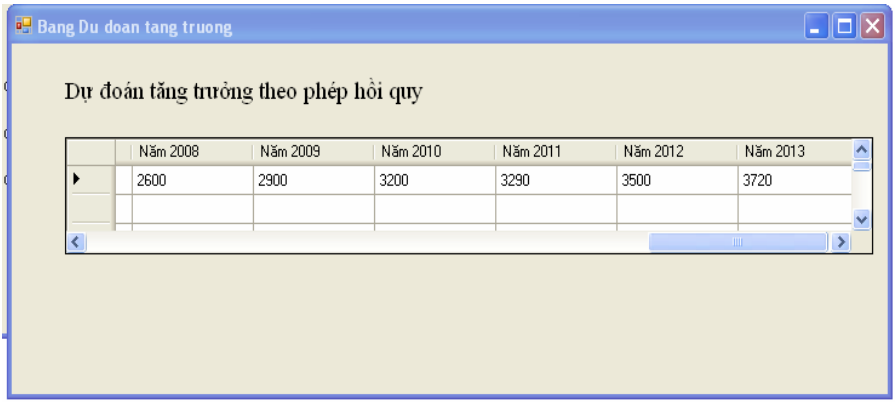



Hình 3.5: Bảng cơ sở dữ liệu Test

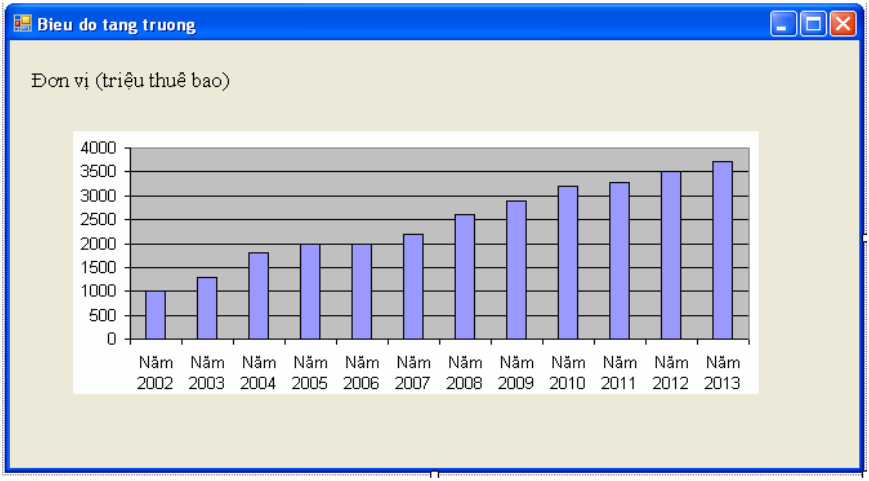
Là tập cơ sở dữ liệu có được sau khi tiến hành kiểm tra trên một tập cơ sở dữ liệu thuê bao khi dùng phương pháp cây quyết định. CSDL này cho phép dự đoán những người có khả năng rời bỏ hay không.

Menu xem dữ liệu gốc cho phép xem tập cơ sở dữ liệu huấn luyện hoặc một một

**Dự đoán tăng trưởng số thuê bao theo phương pháp hồi quy.**




Hình 3.7: Dự đoán tăng trưởng số thuê bao hằng năm




Hình 3.8: Biểu đồ tăng trưởng

## 3.4. ĐÁNH GIÁ

Phần mềm còn hạn chế nhưng đã đưa ra một số kết quả nhất định, cho phép truy cập dữ liệu, xem dữ liệu gốc, đưa ra khả năng dự đoán cho mỗi khách hàng và tăng trưởng hằng năm. Để có có giá trị sử dụng cao cần có đầu tư thời gian nhiều hơn nữa.

# KẾT LUẬN

Đề tài *Ứng dụng Khai phá dữ liệu để dự đoán sự tăng trưởng số thuê bao di động* về cơ bản đã đáp ứng được các yêu cầu đặt ra. Đề tài đã xây dựng được phần mềm có các chức năng khai phá dữ liệu đáp ứng đầu ra của bài toán phục vụ công tác tham mưu, quản lý trong việc quy hoạch phát triển số thuê bao di động.

**Các kết quả đạt được của đề tài:**

• Nắm vững hơn kiến thức về công nghệ: Quy trình khai phá dữ liệu, DotNet, SQL 2005.

• Nâng cao tính làm việc theo nhóm, khả năng tìm kiếm tài liệu, thông tin, các kỹ thuật trên cơ sở dữ liệu, như trích lọc, biến đổi, thu gọn dữ liệu.

• Đã tiến hành thu thập, tổng hợp về các thông tin, quản lý khách hàng sử dụng thuê bao di động, nghiệp vụ xử lý cước.

• Đã thực hiện tốt các giải thuật cây quyết định để phân lớp khách hàng: đưa ra dự đoán.

- Khách hàng tiềm năng.
- Khách hàng rời bỏ.
- Tính ra được số phần trăm thuê bao ảo.
- Ước lượng, điều chỉnh nhu cầu sử dụng và áp dụng đầu tư công nghệ đáp ứng được công nghệ cho mạng di động hoạt động tốt.

• Cho phép người dùng khai thác có thể tra dự đoán tăng trưởng hàng năm đưa ra dưới dạng hai hình thức.

- Dạng biểu đồ
- Dạng bảng

• Đánh giá xu hướng biến đổi sản lượng của các sản phẩm, dịch vụ: mục tiêu của chức năng này là từ thông tin về tình hình sản xuất, kinh doanh các sản phẩm, dịch vụ.

• Cho phép người quản trị cập nhật thông tin một cách nhanh chóng, đơn giản đáp ứng nhu cầu quản lý và khai thác thông tin.

**Những hạn chế của đề tài:**

- Đề tài đã cố gắng thu thập, tổng hợp, phân tích dữ liệu đưa ra các dự đoán khác nhau. Tuy nhiên, do số liệu thu thập là còn ít dựa trên mẫu chưa phản ánh tình hình khách quan, trong khi thực tế lại là một cơ sở dữ liệu rất lớn.

- Các giải thuật chưa phải là giải pháp tối ưu để lựa chọn các mẫu thông tin cần thiết.

- Các số liệu thu thập và phân tích chưa đồng bộ.

Hướng phát triển: Do đề tài triển khai xây dựng cơ sở dữ liệu bản ghi nên tương tác trên cơ sở dữ liệu là rất lớn nên việc cập nhật các dữ liệu rất khó khăn. Cần xây dựng một và quản lý cơ sở dữ liệu tối ưu để thực hiện chức năng dự đoán chính xác và khách quan hơn.