LỜI CẢM ƠN

Tôi xin trân trọng cảm ơn các thầy cô trong Khoa công nghệ thông tin đã tạo điều kiện cho tôi một môi trường học tập tốt đồng thời truyền đạt cho tôi một vốn kiến thức quý báu một tư duy khoa học để phục vụ cho quá trình học tập và công tác của tôi

Tôi xin gửi lời cảm ơn đến các bạn trong lớp Cao học Hệ thống thông tin M13CQIS02-B khóa 2013- 2015 đã giúp đỡ tôi trong suốt thời gian học tập vừa qua.

Đặc biệt, tôi xin được bày tỏ lòng biết ơn sâu sắc đến **TS. HOÀNG XUÂN DẬU** đã tận tình chỉ bảo cho tôi trong suốt quá trình học tập và nghiên cứu, giúp tôi có nhận thức đúng đắn về kiến thức khoa học, tác phong học tập và làm việc.

Cuối cùng, tôi xin được gửi lời cảm ơn tới gia đình, đồng nghiệp, người thân đã động viên, giúp đỡ tôi trong quá trình hoàn thành luận văn.

Hà Nội, Tháng 7 năm 2015

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Lê Văn Trường

MỤC LỤC

DANH MỤC KÝ HIỆU CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
IG	Information Gain	Lượng thông tin thu thêm
NV	Naïve Bayes	Bayes đơn giản
SVM	Support Vector Machine	Máy học véc tơ hỗ trợ
kNN	k-Nearest Neighbor	K hàng xóm gần nhất
MAP	Maximum A Posterior Probability	Phương pháp xác suất hậu nghiệm cực đại
MLE	Maximum Likelihood Estimation	Quy tắc giả thuyết hợp lý nhất
KKT	Karush –Kuhn-Tucker	
PL/SQL	Procedural Language/ Structured Query Language	Ngôn ngữ thủ tục/Ngôn ngữ truy vấn cấu trúc
Java APIs	Java Application Programming Interface	Giao diện lập trình ứng dụng Java
ODM	Oracle Data Mining	Khai phá dữ liệu Oracle
PTTB		Phát triển thuê bao
CSDL		Cơ sở dữ liệu

DANH MỤC CÁC BẢNG

DANH MỤC HÌNH VẼ, ĐỒ THỊ

LỜI MỞ ĐẦU

Trong thời gian qua, lĩnh vực dịch vụ viễn thông của Việt Nam đã đạt được những thành tựu nổi bật với tốc độ tăng trưởng vượt trội so với các ngành dịch vụ khác. Thị trường viễn thông hiện nay bao gồm nhiều loại hình dịch vụ khác nhau. Ngành viễn thông của Việt Nam chỉ thật sự bắt đầu bước vào cạnh tranh từ năm 2003 sau khi một số nhà khai thác mới được cấp phép cung cấp dịch vụ. Cùng với đó, sự cạnh tranh giữa các nhà cung cấp dịch vụ cũng trở nên sôi động và quyết liệt. Bên cạnh các hoạt động cạnh tranh lành mạnh đúng pháp luật, đã xảy ra nhiều hoạt động cạnh tranh gay gắt, thiếu lành mạnh, chèn ép trong hoạt động cung cấp dịch vụ, thậm chí vi phạm luật cạnh tranh của một số doanh nghiệp trong lĩnh vực này.

Trong giai đoạn đổi mới, VNPT Hà Nội luôn là đơn vị đi đầu cả nước trong triển khai và đưa vào sử dụng các công nghệ viễn thông hiện đại với nhiều loại hình dịch vụ mới, thể hiện những thành tựu mới nhất trong công nghệ viễn thông như truyền hình tương tác thế hệ mới MyTV, dịch vụ Internet cáp quang fiberVNN, các dịch vụ truyền dữ liệu, dịch vụ gia tăng trên điện thoại di động ...

Trong môi trường cạnh tranh khốc liệt của thị trường viễn thông , VNPT Hà Nội luôn xác định tầm quan trọng của công tác phân loại và chăm sóc khác hàng nhằm đảm bảo chất lượng dịch vụ và sự hài lòng của khách hàng khi sử dụng dịch vụ.VNPT Hà Nội luôn coi vũ khí để cạnh tranh tốt nhất là chất lượng dịch vụ và phục vụ. Với mục tiêu phân nhóm khách hàng để tiếp thị và bán hàng hiệu quả nhằm tạo ra những gói dịch vụ, những gói cước đa dạng và linh hoạt, phù hợp với nhiều đối tượng khác nhau để từ đó có những động thái chăm sóc khách hàng tốt nhất, khẳng định vị thế thương hiệu của VNPT Hà Nội.

Có một số phương pháp phân lớp được đề xuất sử dụng, tuy nhiên kết quả đạt được còn hạn chế do thiếu các cơ sở lý thuyết vững chắc và đa số thường chỉ được sử dụng để giải quyết một trường hợp cụ thể, như xây dựng một số báo cáo theo yêu cầu của Lãnh đạo. Đề tài " Ứng dụng học máy trong phân loại khách hàng

tại VNPT Hà Nội " nhằm giải quyết bài toán phân lớp khách hàng một cách có hệ thống trên cơ sở lý thuyết vững chắc, đáp ứng yêu cầu quản lý, điều hành.

Việc lựa chọn đề tài trên với các mục đích sau:

Nghiên cứu tổng quan về học máy và các phương pháp học máy.

Nghiên cứu sâu hai thuật toán học máy Naïve Bayes và Support Vector Machine (SVM).

Cài đặt và thử nghiệm hai thuật toán học máy trên bộ dữ liệu khách hàng của VNPT Hà Nội và đánh giá kết quả thu được.

Luận văn sẽ được trình bày với 3 chương chính với nội dung như sau:

Chương 1: Tổng quan về học máy và ứng dụng

Chương I trình bày một cách tổng quan nhất về các khái niệm và thuật toán học máy phổ biến và đưa ra những yêu cầu khái quát nhất về bài toán phân loại khách hàng tại VNPT Hà Nội. Sau khi đã tìm hiểu, so sánh các thuật toán học máy và khả năng ứng dụng của các thuật toán với bài toán phân loại khách hàng, luận văn đã đưa ra lựa chọn về thuật toán cho bài toán.

Chương 2: Xây dựng mô hình phân loại khách hàng dựa trên Naïve Bayes và Support Vector Machine (SVM)

Chương 2 tập trung nghiên cứu về hai thuật toán là Naïve Bayes và Support Vector Machine (SVM) để hiểu rõ việc thực hiện huấn luyện và phân loại đồng thời cũng mô tả mô hình phân loại khách hàng bằng hai thuật toán dựa trên dữ liệu thực tế về khách hàng và thuê bao của VNPT Hà Nội.

Chương 3: Thử nghiệm và kết quả

Chương 3 trình bày mô hình phân loại mà luận văn đã đề xuất ở chương 2 và cách thức cài đặt mô hình này. Tiếp theo là thử nghiệm 2 bộ phân loại Naïve Bayes và SVM trên tập dữ liệu khách hàng và thuê bao của VNPT Hà Nội sử dụng công cụ Oracle Data Mining. Cuối cùng thực hiện đánh giá và so sánh kết quả thử nghiệm của 2 bộ phân loại.

CHƯƠNG I. TỔNG QUAN VỀ HỌC MÁY VÀ ỨNG DỤNG

1.1. Giới thiệu về học máy

Học máy (machine learning) là khả năng của chương trình máy tính sử dụng kinh nghiệm, quan sát, hoặc dữ liệu trong quá khứ để cải thiện công việc của mình trong tương lai. Chẳng hạn, máy tính có thể học cách dự đoán dựa trên các ví dụ, hay học cách tạo ra các hành vi phù hợp dựa trên quan sát trong quá khứ [2].

Một số khái niệm

Mẫu hay ví dụ là tên gọi đối tượng cần phân loại. Chẳng hạn, khi lọc thư rác, mỗi thư gọi là một mẫu hay một ví dụ.

Mẫu thường được mô tả bằng một tập các *thuộc tính*, còn được gọi là *đặc trưng* hay *biến*.

Ví dụ, trong bài toán chẩn đoán bệnh, thuộc tính là những triệu chứng của người bệnh và các tham số khác: cân nặng, huyết áp, v.v.

Nhãn phân loại thể hiện loại của đối tượng mà ta cần dự đoán. Đối với trường hợp phân loại thư rác, nhãn phân loại có thể là "rác" hay "bình thường". Trong giai đoạn học hay còn gọi là giai đoạn huấn luyện, thuật toán học được cung cấp cả nhãn phân loại của mẫu, trong giai đoạn dự đoán, thuật toán chỉ nhận được các mẫu không nhãn và cần xác định nhãn cho những mẫu này.

Kết quả học thường được thể hiện dưới dạng một ánh xạ từ mẫu sang nhãn phân loại. Ánh xạ này được thể hiện dưới dạng một hàm gọi là hàm đích (target function) có dạng $f : X \to C$. Trong đó X là không gian các ví dụ và C là tập các nhãn phân loại khác nhau.

1.2. Phân loại các phương pháp học máy

1.2.1. Học máy có giám sát, bán giám sát và không giám sát

1.2.1.1. Học máy có giám

Học có giám sát (supervised learning) là một kĩ thuật của ngành học máy để xây dựng một hàm (function) từ tập dữ liệu huấn luyện. Dữ liệu huấn luyện bao

gồm các cặp gồm đối tượng đầu vào (thường dạng vec-tơ), và đầu ra mong muốn. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là hồi qui), hay có thể là dự đoán một nhãn phân loại cho một đối tượng đầu vào (gọi là phân loại). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một đối tượng bất kì là đầu vào hợp lệ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hóa dữ liệu sẵn có để dự đoán được những tình huống chưa gặp phải theo một cách hợp lí.

Học có giám sát có thể tạo ra 2 loại mô hình. Phổ biến nhất, học có giám sát tạo ra một mô hình toàn cục (*global model*) để ánh xạ đối tượng đầu vào đến đầu ra mong muốn. Tuy nhiên, trong một số trường hợp, việc ánh xạ được thực hiện dưới dạng một tập các mô hình cục bộ, dựa trên các "hàng xóm" của nó.

Để giải quyết một bài toán học có giám sát (ví dụ: để nhận dạng chữ viết tắt) người ta phải xét nhiều bước khác nhau:

- Xác định loại của tập dữ liệu huấn luyện. Trước khi làm bất cứ điều gì, chúng ta nên quyết định loại dữ liệu nào sẽ được sử dụng làm dùng để huấn luyện. Chẳng hạn, đó có thể là một kí tự viết tay đơn lẻ, toàn bộ một từ viết tay, hay toàn bộ một dòng chữ viết tay.
- Thu thập dữ liệu huấn luyện. Tập dữ liệu huấn luyện cần phù hợp với các hàm chức năng được xây dựng. Vì vậy, cần thiết phải kiểm tra tích thích hợp của dữ liệu đầu vào để được dữ liệu đầu ra tương ứng. Tập dữ liệu huấn luyện có thể được thu thập từ nhiều nguồn khác nhau: từ việc đo được tính toán, từ các tập dữ liệu có sẵn…
- Xác định việc biểu diễn các đặc trưng đầu vào cho hàm chức năng. Sự chính xác của hàm chức năng phụ thuộc lớn vào cách biểu diễn các đối tượng đầu vào. Thông thường, đối tượng đầu vào được chuyển đổi thành một vec-tơ đặc trưng, chứa một số các đặc trưng nhằm mô tả cho đối tượng đó. Số lượng các đặc trưng không nên quá lớn, do sự bùng nổ dữ liệu, nhưng phải đủ lớn để dự đoán chính xác đầu ra. Nếu hàm chức năng mô tả quá chi tiết về đối tượng, thì các dữ liệu đầu ra có

thể bị phân rã thành nhiều nhóm hay nhãn khác nhau, việc này dẫn tới việc khó phân biệt được mối quan hệ giữa các đối tượng hay khó tìm được nhóm (nhãn) chiếm đa số trong tập dữ liệu cũng như việc dự đoán phần tử đại diện cho nhóm, đối với các đối tượng gây nhiễu, chúng có thể được dán nhãn, tuy nhiễn số lượng nhãn quá nhiều, và số nhãn tỉ lệ nghịch với số phần của mỗi nhãn. Ngược lại, hàm chức năng có quá ít mô tả về đối tượng dễ dẫn tới việc dán nhãn đối tượng bị sai hay dễ bỏ xót các đối tượng gây nhiễu. Việc xác định tương đối đúng số lượng đặc tính của phần tử sẽ giảm bớt chi phí khi thực hiện đánh giá kết quả sau huấn luyện cũng như kết quả gặp bộ dữ liệu đầu vào mới.

- Xác định cấu trúc của hàm chức năng cần tìm và giải thuật học tương ứng. Ví dụ, người kĩ sư có thể lựa chọn việc sử dụng mạng nơ-ron nhân tạo hay cây quyết định.

- Hoàn thiện thiết kế. Người thiết kế sẽ chạy giải thuật học từ tập huấn luyện thu thập được. Các tham số của giải thuật học có thể được điều chỉnh bằng cách tối ưu hóa hiệu năng trên một tập con (gọi là tập kiểm chứng -validation set) của tập huấn luyện, hay thông qua kiểm chứng chéo (cross-validation). Sau khi học và điều chỉnh tham số, hiệu năng của giải thuật có thể được đo đạc trên một tập kiểm tra độc lập với tập huấn luyện.

1.2.1.2. Học không giám sát

Học không giám sát (unsupervised learning) là một phương pháp nhằm tìm ra một mô hình mà phù hợp với các tập dữ liệu quan sát. Nó khác biệt với học có giám sát ở chỗ là đầu ra đúng tương ứng cho mỗi đầu vào là không biết trước. Trong học không giám sát, đầu vào là một tập dữ liệu được thu thập. Học không giám sát thường đối xử với các đối tượng đầu vào như là một tập các biến ngẫu nhiên. Sau đó, một mô hình mật độ kết hợp sẽ được xây dựng cho tập dữ liệu đó.

Học không giám sát có thể được dùng kết hợp với suy diễn Bayes (*Bayesian inference*) để cho ra xác suất có điều kiện cho bất kì biến ngẫu nhiên nào khi biết trước các biến khác.

Học không giám sát cũng hữu ích cho việc nén dữ liệu: về cơ bản, mọi giải thuật nén dữ liệu hoặc là dựa vào một phân bố xác suất trên một tập đầu vào một cách tường minh hay không tường minh.

1.2.1.3. Học bán giám sát

Học bán giám sát (semi-supervised learning) là dạng kết hợp giữa học có giám sát và học không giám sát, trong đó, nó kết hợp các ví dụ có gắn nhãn và không gắn nhãn để sinh một hàm hoặc một bộ phân loại thích hợp.

1.2.2. Một số thuật toán học máy thông dụng

Trong mục này Luận văn trình bày chi tiết về các thuật toán học máy thông dụng như: Cây quyết định và k-láng giềng gần nhất, còn các thuật toán học máy Naïve Bayes và SVM sẽ được trình bày chi tiết trong chương II.

1.2.2.1. Cây quyết định [2]

1.2.2.1.1. Tổng quan về thuật toán cây quyết định

Chúng ta có thể định nghĩa cây quyết định có các tính chất sau:

- Mỗi nút trong (internal node) biểu diễn một thuộc tính cần kiểm tra giá trị (anattribute to be tested) đối với các các tập thuộc tính.
- Nút lá (leaf node) hay còn gọi là nút trả lời biểu thị cho một lớp các trường hợp mà nhãn của nó là tên của lớp, nó biểu diễn một lớp (a classififcation).
- Nút nhánh (branch) từ một nút sẽ tương ứng với một giá trị có thể của thuộc tính gắn với nút đó.
- Nhãn (lable) của nút này là tên của thuộc tính và có một nhánh nối nút này đến các cây con ứng với mỗi kết quả có thể có phép thử. Nhãn của nhánh này là các giá trị của thuộc tính đó. Nút trên cùng gọi là nút gốc.

Hình 1.1. Mô tả chung về cây quyết định

1.2.2.1.2. Thiết kế cây quyết định

a) Xử lý dữ liệu

Công việc cụ thể của bước tiền xử lý dữ liệu gồm các công việc:

- Lọc thuộc tính(Filtering Attributes)
- Lọc các mẫu(Filtering samples)
- Lọc các mẫu (instances, patterns)
- Chuyển đổi dữ liệu(Transformation)
- Rời rạc hóa dữ liệu(Discretization)

b) Tạo cây

Cây quyết định được tạo thành bằng cách lần lượt chia (theo phương pháp đệ quy) một tập dữ liệu thành các tập dữ liệu con, mỗi tập con được tạo thành từ các phần tử của cùng một lớp. Các nút (không phải là nút lá) là các điểm phân nhánh của cây. Việc phân nhánh tại các nút có thể dựa trên việc kiểm tra một hay nhiều thuộc tính để xác định việc phân chia dữ liệu.

c) Tiêu chuẩn tách

Chúng ta mong muốn chọn thuộc tính sao cho việc phân lớp tập mẫu là tốt nhất. Như vậy chúng ta cần phải có một tiêu chuẩn để đánh giá vấn đề này. Có rất nhiều tiêu chuẩn được đánh giá được sử dụng đó là: Lượng thông tin thu thêm IG (Information Gain), thuật toán ID3 của John Ross Quilan.

d) Tiêu chuẩn dừng

Chúng ta tập trung một số tiêu chuẩn dừng chung nhất được sử dụng trong cây quyết định. Tiêu chuẩn dừng truyền thống sử dụng các tập kiểm tra. Chúng ta có thể thay ngưỡng như là giảm nhiễu, số các mẫu trong một nút, tỉ lệ các mẫu trong nút, hay chiều sâu của cây.

e) Tỉa cây

Sau giai đoạn tạo cây chúng ta có thể dùng phương pháp "Độ dài mô tả ngắn nhất" (Minimum Description Length) hay giá trị tối thiểu của IG để tỉa cây (chúng ta có thể chọn giá trị tối thiểu của IG trong giai đoạn tạo cây đủ nhỏ để cho cây phát triển tương đối sâu, sau đó lại nâng giá trị này lên để tỉa cây).

f) Các bước tổng quát để xây dựng cây quyết định

Quá trình xây dựng một cây quyết định cụ thể bắt đầu bằng một nút rỗng bao gồm toàn bộ các đối tượng huấn luyện và làm như sau :

- Nếu tại nút hiện thời, tất cả các đối tượng huấn luyện đều thuộc vào một lớp nào đó thì nút này chính là nút lá có tên là nhãn lớp chung của các đối tượng.
- Trường hợp ngược lại, sử dụng một độ đo, chọn thuộc tính điều kiện phân chia tốt nhất tập mẫu huấn luyện có tại nút.
- Tạo một lượng nút con của nút hiện thời bằng số các giá trị khác nhau của thuộc tính được chọn. Gán cho mỗi nhánh từ nút cha đến nút con một giá trịcủa thuộc tính rồi phân chia các các đối tượng huấn luyện vào các nút con tương ứng.
- Nút con K được gọi là thuần nhất, trở thành lá, nếu tất cả các đối tượng mẫu tại đó đều thuộc vào cùng một lớp.
- Lặp lại các bước 1 - 3 đối với mỗi nút chưa thuần nhất.

1.2.2.1.3 Thuật toán xây dựng cây quyết định dựa vào Entropy

Tiêu chí để đánh giá tìm điểm chia là rất quan trọng, chúng được xem là một tiêu chuẩn "heuristic" để phân chia dữ liệu. Ý tưởng chính trong việc đưa ra các tiêu chí trên là làm sao cho các tập con được phân chia càng trở nên "trong suốt" (tất cả các bộ thuộc về cùng một nhãn) càng tốt. Thuật toán dùng độ đo lượng thông tin thu

thêm (Information Gain - IG) để xác định điểm chia. Độ đo này dựa trên cơ sở lý thuyết thông tin của nhà toán học Claude Shannon, độ đo này được xác như sau:

Xét bảng quyết định DT = (U, C ∪{d}), số giá trị (nhãn lớp) có thể của d là k. Khi đó Entropy của tập các đối tượng trong DT được định nghĩa bởi:

$$Entropy(U) = -\sum_{i=0}^{k} p_i \log_2 p_i$$

Trong đó p_i là tỉ lệ các đối tượng trong DT mang nhãn lớp i.

Ý nghĩa của đại lượng Entropy trong lĩnh vực lý thuyết công nghệ thông tin: Entropy của tập U chỉ ra số lượng bít cần thiết để mã hóa lớp của một phần tử được lấy ra ngẫu nhiên từ tập U. Lượng thông tin thu thêm (Information Gain- IG) là lượng Entropy còn lại khi tập các đối tượng trong DT được phân hoạch theo một thuộc tính điều kiện c nào đó. IG xác định theo công thức:

$$IG(U,c) = Entropy(\mathrm{U}) - \sum_{v \in V_c} \frac{|U_v|}{|U|} Entropy(U_v)$$

Trong đó V_c là tập các giá trị của thuộc tính c, U_v là tập các đối tượng trong DT có giá trị thuộc tính c bằng v. Giá trị IG(U, c) được sử dụng làm độ đo lựa chọn thuộc tính phân chia dữ liệu tại mỗi nút trong thuật toán xây dựng cây quyết định ID3. Thuộc tính được chọn là thuộc tính cho lượng thông tin thu thêm lớn nhất. Ý nghĩa của đại lượng IG trong lĩnh vực lý thuyết công nghệ thông tin: IG của tập S chỉ ra số lượng bít giảm đối với việc mã hóa lớp của một phần tức được lấy ra ngẫu nhiên từ tập U.

1.2.2.2. K hàng xóm gần nhất [5][2]

K-hàng xóm gần nhất (k-nearest neighbors, viết tắt là k-NN) là phương pháp tiêu biểu nhất của học dựa trên ví dụ. Nguyên tắc của phương pháp này là đặc điểm của mẫu được quyết định dựa trên đặc điểm của k mẫu giống mẫu đang xét nhất. Ví dụ, muốn xác định nhãn phân loại, ta tìm k mẫu gần nhất và xem những mẫu này mang nhãn gì.

Phương pháp k-NN thường làm việc với dữ liệu trong đó các thuộc tính được cho dưới dạng vec tơ các số thực. Như vậy, mỗi mẫu tương ứng với một điểm trong

không gian Ơclit. Giả sử mẫu x có giá trị thuộc tính là $< a_1(x), a_2(x), ..., a_n(x) >$. Để xác định các mẫu giống x, cần có độ đo khoảng cách giữa các các mẫu. Vì mẫu trương ứng với điểm trong không gian, khoảng cách Ơclit thường được dùng cho mục đích này. Khoảng cách Ơ clit giữa hai mẫu x_i và x_j được tính như sau:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (a_l(x_i) - a_l(x_j))^2}$$

Với khoảng cách $d(x_i, x_j)$ vừa được định nghĩa, phương pháp k-NN cho hai trường hợp: phân loại và hồi quy(regrestion) được thực hiện như sau.

- Phân loại

Mỗi mẫu x có thể nhãn phân loại với f(x) với f(x) nhận một giá trị trong tập hữu hạn các phân loại *C*. Thuật toán k-NN cho phân loại được cho như sau.

Thuật toán:

Giai đoạn học (huấn luyện)

Lưu các mẫu huấn luyện có dạng <x,f(x)> vào cơ sở dữ liệu

Giai đoạn phân loại

Đầu vào: tham số k

Với mẫu x cần phân loại:

a) Tính khoảng cách $d(x, x_i)$ từ x tới tất cả mẫu x_i trong cơ sở dữ liệu

b) Tìm k mẫu có $d(x, x_i)$ nhỏ nhất, giả sử k mẫu đó là $\{x_1, x_2, ..., x_k\}$.

c) Xác định nhãn phân loại $f'(x)$ là nhãn chiếm đa số trong tập $\{x_1, x_2, ..., x_k\}$

- Hồi quy(Regression)

Mỗi mẫu x có thể nhãn phân loại $f(x)$ với $f(x)$ là một số thực. Thuật toán k-NN ở trên có thể thay đổi dễ dàng cho bài toán hồi quy này bằng cách thay bước đánh số 3 trong thuật toán như sau:

$$f'(x) = \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

Thuật toán k-NN có một tham số đầu vào là k: số hàng xóm được dùng để quyết định nhãn cho mẫu đang xét. Nếu k=1, giá trị hàm $f'(x)$ được chọn bằng giá trị hàm f của mẫu gần nhất. Thông thường k=1 không cho kết quả tốt do hàng xóm gần nhất có ảnh hưởng quyết định tới giá trị $f'(x)$. Trong trường hợp hàng xóm gần nhất là nhiễu sẽ khiến kết quả bị sai. Nhiều nghiên cứu cho thấy giá trị k trong khoảng từ 5 đến 10 là phù hợp. Để xác định giá trị cụ thể của k có thể sử dụng phương pháp kiểm tra chéo đã trình bày ở phần tỉa cây. Giá trị k cho độ chính xác khi kiểm tra chéo tốt nhất sẽ được lựa chọn cho thuật toán.

1.3. Ứng dụng của học máy

- **Ứng dụng dự đoán tội phạm của IBM**

Một trong những ứng dụng nổi bật về phân lớp dữ liệu phải kể đến là phương pháp dự đoán tội phạm của IBM. Sự phân lớp của họ dựa trên một lượng lớn các sự kiện diễn ra ở thành phố New York mỗi ngày (thời tiết, sự kiện thể thao chính, thời gian trong ngày, thời gian trong năm, vv…) và sau đó xác định ở đâu trong thành phố sẽ có tội phạm. Mục đích cuối cùng là xác định xem liệu có tội phạm trong một khu vực xác định nào đó không. Ví dụ (có tội phạm ở Lower Manhattan = có/ không = hình tròn/ hình vuông). Sau khi thu thập dữ liệu của vài năm họ phân lớp và bắt đầu sử dụng để dự đoán khu vực nào sẽ có vụ phạm tội trong một ngày bất kỳ. Vì vậy, ở đầu ngày hôm đó họ sẽ cho các dữ liệu vào máy tính ví dụ (ngày hôm nay nắng, Thứ Hai, có một sự kiện của người dân New York tại đâu đó, vv…) và máy tính sẽ tìm các ngày trong quá khứ có dữ liệu gần tương tự, và xác định xem đâu là những điểm nóng về tội phạm của ngày hôm đó. Xe cứu thương và xe cảnh sát sẽ tuần tra khu vực đó và đợi tội phạm xuất hiện. Phương pháp này theo báo cáo đã giúp giảm số lần ứng phó khẩn cấp ở New York.

- **Ứng dụng Nhận diện khuôn mặt trong máy ảnh kỹ thuật số**

Một ứng dụng phổ biến của phân lớp dữ liệu là nhận dạng khuôn mặt. Máy tính lấy hàng trăm hình ảnh của mọi người và chỉ ra đâu là vị trí khuôn mặt trong bức hình (nó được thực hiện bằng cách vẽ một hộp quanh mỗi khuôn mặt, trong mỗi bức hình). Máy tính nhận dữ liệu là vị trí và màu của các pixel và mối quan hệ giữa

chúng với nhau, và sau đó phân thành lớp có khuôn mặt và không có khuôn mặt trong hình. Máy tính phải được huấn luyện bằng cách cho biết khuôn mặt sẽ hiển thị như thế nào trong bức hình hàng ngàn lần. Từ đó nó có thể nhận diện được có khuôn mặt trong hình hay không, và nó ở đâu (có mặt /không có mặt = hình tròn và hình vuông). Công nghệ này hiệu quả , chi phí tính toán thấp và được dùng trong máy ảnh kỹ thuật số.

- **Ứng dụng lọc thư rác tự động**

Khi một bức thư được gửi đến Mail server exchange, nhờ chức năng bẫy sự kiện Incoming mail của SMTP thì bức thư đó được đưa đến Bộ phân loại Tiếng Anh, tiếng Việt hoặc đưa thẳng đến Bộ phân loại thư rác (phụ thuộc vào tùy chọn của người dùng).

Giả sử sau khi đưa vào bộ phân loại Tiếng Anh, tiếng Việt, bức thư được chuyển cho bộ phân loại thư rác. Tại đây, nhờ quá trình tính toán theo Naive Bayes, nó sẽ được gán nhãn là thư rác [Possible Spam] hoặc không gán nhãn nếu được xác định là thư thường. Sau đó, thư được gửi đến Exchange server nhờ dịch vụ SMTP. Cuối cùng, bức thư đã sẵn sàng cho Mail Client lấy về qua giao thức POP3.

1.4. Giới thiệu bài toán phân loại khách hàng tại VNPT Hà Nội

1.4.1 . Mô tả bài toán

1.4.1.1. Một số định nghĩa

a) Khái niệm về khách hàng.

Theo cách hiểu chung nhất của các nhà kinh tế, khách hàng là tất cả những người (cá nhân, tập thể hay tổ chức) có nhu cầu và thực hiện trực tiếp hoặc gián tiếp việc giao dịch mua bán hàng hoá hay dịch vụ với các cơ sở sản xuất, doanh nghiệp, các cửa hàng…

Đặc điểm khách hàng sử dụng dịch vụ viễn thông

- Khách hàng sử dụng dịch vụ viễn thông là tất cả mọi người dân thuộc các tầng lớp trong xã hội không phân biệt tuổi tác, giới tính, thu nhập, nghề nghiệp, tôn giáo, lối sống... dẫn đến sở thích và thói quen tiêu dùng sẽ rất đa dạng khác nhau.

- Khách hàng trên thị trường dịch vụ viễn thông (bao gồm cả cá nhân, tổ chức trực tiếp sử dụng dịch vụ như người tiêu dùng cuối cùng) rất khác nhau về qui mô cũng như nhu cầu và mong muốn, dẫn đến sự chênh lệch giữa khách hàng là một cá nhân và khách hàng là một tổ chức về mức độ tiêu dùng.

- Thị trường viễn thông tại Việt Nam cũng rất khác nhau theo vùng địa lý hành chính. Trong khi ở những vùng đồng bằng, thành phố lớn, nơi tập trung đông cơ quan, dân cư có mức thu nhập khá, nhu cầu phát sinh nhiều và đa dạng với yêu cầu chất lượng cao, thì tại các vùng cao, vùng kinh tế khó khăn, nhu cầu sử dụng rất khiêm tốn, hầu như khách hàng chỉ có nhu cầu sử dụng các dịch vụ cơ bản do khả năng kinh tế hạn chế.

- Trong hoạt động mua bán, khách hàng có vai trò quyết định bởi họ mang nhu cầu (đối tượng lao động) đến bất cứ lúc nào và tham gia trực tiếp vào quá trình sản xuất tạo ra dịch vụ. Hành vi mua của khách hàng là một yếu tố đặc biệt chi phối năng suất của dịch vụ bởi nó tác động trực tiếp đến việc cung cấp và tiêu dùng dịch vụ. Khách hàng sử dụng dịch vụ viễn thông cũng không đồng đều về thời gian như giờ trong ngày, ngày trong tuần, tuần trong tháng và tháng trong năm.

- Khi sử dụng dịch vụ viễn thông, khách hàng không đơn thuần chỉ sử dụng một dịch vụ mà còn có thể muốn mua nhiều hơn một sản phẩm hay dịch vụ và luôn muốn được đối xử tốt.

- Khách hàng sử dụng dịch vụ viễn thông khá nhạy cảm và dễ bị chi phối bởi môi trường xung quanh trong quá trình sử dụng dịch vụ.

b) Phân loại khách hàng.

Phân loại khách hàng là sắp xếp các khách hàng có những yếu tố khác nhau vào các nhóm khác nhau theo một tiêu chí chuẩn, mà tiêu chí đó được đánh giá là cơ sở cho việc điều tiết mối quan hệ giữa nhà cung cấp dịch vụ với khách hàng. Có nhiều tiêu chí để phân loại khách hàng. Tuy nhiên, dù phân loại khách hàng theo tiêu chí nào cũng đều với mục đích cuối cùng là hướng tới thị trường người tiêu dùng. Trong công tác Chăm sóc khách hàng(CSKH), căn cứ vào đặc tính của sản phẩm dịch vụ và mục đích cần hướng tới mà có thể lựa chọn cách phân loại tương

ứng để thực hiện các hoạt động CSKH một cách linh hoạt, sáng tạo, hiệu quả phù hợp với đặc điểm của từng đối tượng.

Một số tiêu chí chủ yếu thường được sử dụng để phân loại khách hàng:

❖ Căn cứ vào vị trí địa lý:

Vị trí địa lý cư trú khách hàng rất quan trọng trong việc quyết định phân loại các khách hàng đang sử dụng dịch vụ của đơn vị trên thị trường. Theo tiêu chí phân loại này, khách hàng được chia thành các loại như sau:

- Khách hàng tại thành phố, đô thị, khu công nghiệp;
- Khách hàng tại khu vực huyện thị.

❖ Căn cứ vào yếu tố tâm lý:

Nắm bắt tâm lý và mong muốn của khách hàng luôn là yếu tố quan trọng hàng đầu trong bất cứ chiến lược CSKH nào. Tiêu chí tâm lý liên quan đến tính cách và cách cư xử của khách hàng cũng như có ảnh hưởng tới sức mua hàng hóa của từng nhóm khách hàng. Các yếu tố quan trọng về tâm lý thông thường bao gồm:

- Những tác động đến thói quen mua hàng, ví dụ như áp lực của những người cùng địa vị hay trình độ học vấn.
- Sự ưa chuộng các sản phẩm có các thuộc tính, đặc điểm khác biệt so với các sản phẩm tương đương.
- Thông thường có khá nhiều khách hàng có những phản ứng không giống nhau khi nhận được cùng một sự phục vụ. Khi đánh giá dịch vụ CSKH, người tiêu dùng có rất nhiều sự mong đợi mang tính cách cá nhân, cụ thể và rất đặc trưng về sản phẩm hay dịch vụ của một nhãn hiệu cụ thể.
- Sự trung thành với thương hiệu quen và các lợi ích khách hàng tìm được ở sản phẩm hay dịch vụ của nhà cung cấp.

❖ Căn cứ vào đặc điểm phục vụ:

Theo tiêu chí này, các đối tượng khách hàng của doanh nghiệp thường được chia thành 2 nhóm : Khách hàng bên ngoài và khách hàng bên trong.

Khách hàng bên ngoài: Họ là những người trực tiếp trả tiền, người quyết định mua, người sử dụng, người được hưởng quyền lợi từ việc sử dụng các sản

phẩm, dịch vụ. Như vậy có thể nói khách hàng bên ngoài bao gồm những đối tượng sau :

- Người sử dụng : các cá nhân hoặc tổ chức thực sự sử dụng sản phẩm, dịch vụ .
- Người mua : là những người thu thập thông tin về sản phẩm, lựa chọn, ra quyết định mua, trả tiền
- Người thụ hưởng : các cá nhân hoặc tổ chức được hưởng lợi (hoặc bị ảnh hưởng bởi) từ việc sử dụng sản phẩm hoặc dịch vụ

Khách hàng bên trong: Có nhiều bộ phận không hoặc rất ít tiếp xúc với khách hàng nhưng lại đóng vai trò hết sức quan trọng trong việc hình thành sản phẩm dịch vụ phục vụ khách hàng. Đó là những người phải báo cáo (cấp trên), cần chỉ thị (cấp dưới) hay những người đồng nghiệp cần sự hợp tác của bạn. Họ được coi là các khách hàng bên trong (nội bộ) của doanh nghiệp.

❖ Căn cứ vào hành vi của khách hàng.

Phân loại theo hành vi là phương pháp phân loại khách hàng dựa trên cách mà khách hàng phản ứng, sử dụng hay biết về sản phẩm . Phân loại theo hành vi là chủ đề thu hút sự quan tâm và nghiên cứu của các đơn vị cung cấp dịch vụ vì hành vi của khách hàng rất phức tạp và luôn thay đổi theo thời gian.

Sự phức tạp của quyết định sử dụng dịch vụ là do bị ảnh hưởng bởi rất nhiều các yếu tố xung quanh. Vì vậy, quyết định sử dụng dịch vụ sẽ ảnh hưởng đến cách chọn lựa hay hành vi và đó chính là cơ sở mà phân loại theo hành vi hướng tới với mục tiêu là đưa ra các chiến lược tiếp thị và chăm sóc khách hàng phù hợp nhất. Phân loại theo hành vi bao gồm các loại sau: Phân loại theo mức độ hài lòng của khách hàng với dịch vụ, phân loại theo mức độ trung thành, phân loại theo mức độ yêu cầu và sử dụng các dịch vụ.

Một số chỉ tiêu của VNPT Hà Nội liên quan đến phân loại khách hàng căn cứ vào hành vi của khách hàng.

Chỉ tiêu tốc độ tăng trưởng thuê bao (TDTT):

TDTT=((Số TB phát triển mới – Số TB rời mạng)/ Tổng số TB năm trước) x100%.

Tỷ lệ thuê bao rời mạng.

TLRM = (Số lượng TB rời mạng /Số lượng thuê bao mới) x 100%.

Theo các tiêu chí của VNPT Hà Nội liên quan đến phân loại khách hàng , có thể thấy rằng việc phân loại khách hàng phụ thuộc rất lớn vào tiêu chí hành vi của khách hàng, xác định tình trạng phát triển mới hay tháo hủy của thuê bao. Khách hàng của VNPT Hà Nội chính là chủ thuê bao nên ta có thể quy bài toán phân loại khách hàng dựa trên hành vi của khách hàng về bài toán phân loại thuê bao.

1.4.1.2. Mô tả bài toán

Hiện tại, công việc của bộ phận kinh doanh, phát triển thị trường, chăm sóc khách hàng của VNPT Hà Nội nói chung và Trung tâm Kinh doanh nói riêng vẫn phải thực hiện đưa ra dự báo và các chỉ tiêu kinh doanh, phát triển thuê bao, đưa ra chỉ tiêu về tăng trưởng thuê bao và doanh thu. Đây là một công việc rất khó khăn khi hiện tại vẫn dựa vào suy đoán trên cơ sở tổng hợp số liệu mà không có một phương pháp cố định và có tính khoa học.

Mục tiêu của bài toán là xây dựng mô hình phân loại đối tượng khách hàng và dự đoán hành vi của khách hàng có nguy cơ thực hiện tháo hủy dịch vụ để xây dựng các chính sách ưu đãi mới cho các nhóm đối tượng này. Để giải quyết bài toán, luận văn sẽ quy bài toán phân loại đối tượng khách hàng về bài toán phân loại đối tượng thuê bao vì mối liên quan chặt chẽ giữa hai đối tượng này.

Trong quá trình giải quyết bài toán sẽ đưa ra được mô hình phân loại, dự đoán được ảnh hưởng của các thuộc tính tới việc phân loại, đưa ra danh sách các thuê bao có khả năng sẽ thực hiện tháo hủy từ dữ liệu các thuê bao đang thực hiện tạm dừng.

Nội dung chi tiết của bài toán sẽ được mô tả chi tiết ở Chương II.

1.4.2. Lựa chọn thuật toán học máy

Qua quá trình tìm hiểu các thuật toán và nội dung bài toán phân loại khách hàng đặt ra, luận văn tập trung nghiên cứu và áp dụng hai thuật toán học máy Naïve Bayes và SVM vì các lý do sau:

Phương pháp Bayes có thể phân loại cho hai và nhiều hơn hai lớp, được xem có nhiều ưu điểm nhất vì nó đã đạt được mục tiêu về mặt lý thuyết cho bài toán

phân loại của luận văn. Một ưu điểm nổi bật của phương pháp này là tính được xác suất sai lầm trong phân loại mà nó được gọi là sai số Bayes. Sai số Bayes đã được chứng minh là xác suất sai lầm nhỏ nhất trong bài toán phân loại [3].

Support Vector Machine(SVM) là giải thuật phân lớp có hiệu quả cao và được áp dụng nhiều trong lĩnh vực khai phá dữ liệu với nhiều đặc điểm như : giải quyết được bài toán dữ liệu có số chiều lớn, giải quyết vấn đề overfitting (dữ liệu có nhiều và tách rời nhóm hoặc dữ liệu huấn luyện quá ít), có hiệu suất tổng hợp tốt và hiệu suất tính toán cao [4][7].

1.5 . Kết chương

Chương I đã trình bày một cách tổng quan nhất về các khái niệm về học máy và thuật toán học máy phổ biến , các ứng dụng của học máy. Chương này cũng giới thiệu về bài toàn phân loại khách hàng của VNPT Hà Nội, đồng thời cũng đưa ra những yêu cầu khái quát nhất về bài toán phân loại khách hàng tại VNPT Hà Nội. Sau khi đã tìm hiểu, so sánh các thuật toán học máy và khả năng ứng dụng của các thuật toán với bài toán phân loại khách hàng, luận văn đã đưa ra lựa chọn hai thuật toán Naïve Bayes và Support Vector Machine để giải quyết các vấn đề đặt ra cho bài toán.

CHƯƠNG II: XÂY DỰNG MÔ HÌNH PHÂN LOẠI KHÁCH HÀNG DỰA TRÊN NAÏVE BAYES VÀ SUPPORT VECTOR MACHINE (SVM)

2.1. Thuật toán Naïve Bayes [5][9]

2.1.1. Định lý Bayes

Trong phương pháp học có giám sát, với tập dữ liệu huấn luyện *D* quan sát được, ta cần tìm một giả thuyết tốt nhất trong không gian giả thuyết *H* nào đó, dựa trên một tập dữ liệu huấn luyện D đã cho. Khi có thêm tri thức về xác suất tiền nghiệm của các giả thuyết trong *H* và xác suất xuất hiện dữ liệu quan sát được trong từng giả thuyết, phương pháp tìm giả thuyết có xác suất hậu nghiệm cực đại dựa trên định lý Bayes là tiếp cận thông dụng.

❖ **Phát biểu bài toán**

Xét không gian giả thuyết *H*. Ký hiệu *P(h)* là xác suất giả thuyết *h* đúng khi chưa có tập dữ liệu quan sát. *P(h)* được gọi là xác suất tiền nghiệm của *h*, nó phản ánh tri thức chung về khả năng *h* là đúng. Khi không biết được *P(h)*, có thể giả thiết $P(h_i)=P(h_j)$ với mọi mọi h_i,h_j khác nhau.

- Với mỗi tập dữ liệu *D*, ta ký hiệu:
- *P(D)* là xác suất quan sát được dữ liệu tập này (tức là xác suất của *D* không cho trước tri thức về các giả thuyết)
- *P(D|h)* là xác suất quan sát được *D* khi *h* đúng, (*P(x|y)* là xác suất có *x* với điều kiện đã có *y*) *P(h|D)* là xác suất của giả thuyết *h* đúng khi cho trước tập dữ liệu huấn luyện được quan sát D (ta gọi là *xác suất hậu nghiệm* của giả thuyết *h*).

Giả thuyết tốt nhất cần tìm là giả thuyết có *xác suất hậu nghiệm cực đại* trong *H*. Công thức Bayes để tính xác suất hậu nghiệm P(*h*|D dựa vào P(*h*), P(D) và P(D|*h*) là cơ sở của phương pháp này nên nó còn được gọi là *phương pháp học Bayes*.

Công thức Bayes: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Ví dụ về định lý Bayes.

Giả sử chúng ta có tập dữ liệu sau (dự đoán 1 người có chơi tennis)?

Bảng 2.1. Số liệu dự đoán người chơi tennis

Ngày	Ngoài trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
N1	Nắng	Nóng	Cao	Yếu	Không
N2	Nắng	Nóng	Cao	Mạnh	Không
N3	Âm u	Nóng	Cao	Yếu	Có
N4	Mưa	Bình thường	Cao	Yếu	Có
N5	Mưa	Mát mẻ	Bình thường	Yếu	Có
N6	Mưa	Mát mẻ	Bình thường	Mạnh	Không
N7	Âm u	Mát mẻ	Bình thường	Mạnh	Có
N8	Nắng	Bình thường	Cao	Yếu	Không
N9	Nắng	Mát mẻ	Bình thường	Yếu	Có
N10	Mưa	Bình thường	Bình thường	Yếu	Có
N11	Nắng	Bình thường	Bình thường	Mạnh	Có
N12	Âm u	Bình thường	Cao	Mạnh	Có

- Dữ liệu D. *Ngoài trời là Nắng* và *Gió là Mạnh.*
- Giả thiết (phân loại) *h*. Anh ta chơi tennis.
- Xác xuất tiền nghiệm P(*h*). Xác suất rằng anh ta chơi tennis (bất kể Ngoài trời như thế nào và Gió ra sao).
- Xác suất tiền nghiệm P(D). Xác suất rằng *Ngoài trời* là *nắng* và *Gió* là M*ạnh.*

P(D|*h*). Xác suất *Ngoài trời* là *nắng* và *Gió* là *mạnh*, nếu biết rằng anh ta chơi tennis.

P(*h*|D). Xác suất anh ta chơi tennis, nếu biết rằng *Ngoài trời* là *nắng* và *Gió* là *mạnh.*

2.1.2. Phương pháp xác suất hậu nghiệm cực đại (Maximum A Posterior Probability - MAP)

Với tập dữ liệu quan sát được *D* và các phân bố xác suất *P(h), P(D)* và *P(D/h)* đã biết, ta xác định các xác suất hậu nghiệm cho các giả thuyết *h* trong *H* nhờ dùng định lý Bayes. Lời giải là giả thuyết h_{MAP} thuộc *H* có xác suất hậu nghiệm lớn nhất:

$$h_{MAP} = \arg\max\{P(h \mid D) : h \in H\} = \arg\max\left\{\frac{P(D \mid h)P(h)}{P(D)} : h \in H\right\}$$
$$= \arg\max\{P(h \mid D)P(h) : h \in H\}$$

(lưu ý rằng trong bước tính toán cuối trên ta bỏ *P(D)* vì nó độc lập với *h*).

Ví dụ MAP: Tập H bao gồm 2 giả thiết (có thể)

h_1 : Anh ta chơi tennis

h_2 : Anh ta không chơi tennis

Tính giá trị của 2 xác xuất có điều kiện: $P(h_1 \mid D)$, $P(h_2 \mid D)$

Giả thiết có thể nhất $h_{MAP} = h_1$ nếu $P(h_1 \mid D) \geq P(h_2 \mid D)$; ngược lại thì $h_{MAP} = h_2$.

Bởi vì $P(D) = P(D \mid h_1) + P(D \mid h_2)$ là như nhau đối với cả 2 giả thiết h_1 và h_2, nên có thể bỏ qua đại lượng $P(D)$. Vì vậy, cần tính: $P(D \mid h_1).P(h_1)$ và $P(D \mid h_2).$ $P(h_2)$, và đưa ra quyết định tương ứng .

- Nếu $P(D \mid h_1).P(h_1) \geq P(D \mid h_2).P(h_2)$, thì kết luận anh ta chơi tennis
- Ngược lại, thì kết luận là anh ta không chơi tennis

2.1.3. Quy tắc giả thuyết hợp lý nhất (Maximum Likelhood Estimation - MLE)

Khi không có thông tin về xác suất đúng của các giả thuyết trong *H*, ta giả thiết rằng mọi giả thuyết *h* thuộc *H* có cùng xác suất tiền nghiệm:

$P(h_i) = P(h_j)\ \forall\ h_i, h_j \in H$.

Lưu ý rằng *P(D/h)* biểu thị khả năng xuất hiện tập *D* với *h*, lời giải là giả thuyết *h* làm cực đại *P(D/h),* trong trường hợp này gọi *giả thuyết có khả năng nhất* (Maximum Likelhood) hay ***hợp lý nhất*** và được ký hiệu là h_{ML}:

$$h_{ML} = \arg\max\{P(D/h) : h \in H\}$$

Phương pháp này được gọi là phương pháp hợp lý nhất/có khả năng nhất. Trước khi xét các ứng dụng thực tế, ta xét bài toán học khái niệm như là một ứng dụng lý thuyết của phương pháp này.

Ví dụ MLE: Tập H bao gồm 2 giả thiết có thể

h_1: Anh ta chơi tennis

h_2: Anh ta không chơi tennis

D: Tập dữ liệu (các ngày) mà trong đó thuộc tính Ngoài trời có giá trị Nắng và thuộc tính Gió có giá trị Mạnh

Tính 2 giá trị khả năng xảy ra (likelihood values) của dữ liệu D đối với 2 giả thiết: $P(D|h_1)$ và $P(D|h_2)$

- *P(Ngoài trời=nắng, Gió=mạnh|h1)= 1/8*
- *P(Ngoài trời=nắng,Gió=mạnh|h2)= 1/4*

Giả thiết MLE $h_{MLE} = h_1$ nếu $P(D|h_1) \geq P(D|h_2)$; và ngược lại thì $h_{MLE} = h_2$

→ vì *P(Ngoài trời=nắng, Gió=mạnh|h1)< P(Ngoài trời=nắng,Gió=mạnh|h2)*, hệ thống kết luận rằng: **Anh ta sẽ không chơi tennis!.**

2.1.4. Phân lớp Bayes đơn giản (Naïve Bayes)

Biểu diễn bài toán phân loại (classification problem)

+ Một tập học D_*train*, trong đó mỗi ví dụ học x được biểu diễn là một vectơ n chiều: $(x_1, x_2, ..., x_n)$

+ Một tập xác định các nhãn lớp: $C = (c_1, c_2, ..., c_m)$

+ Với một ví dụ mới z, thì z sẽ được phân vào lớp nào?

Mục tiêu: Xác định phân lớp có thể (phù hợp) nhất đối với z

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid z)$$

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid z_1, z_2, ..., z_n)$$

$$c_{MAP} = \arg\max_{c_i \in C} \frac{P(c_i)\, P(z_1, z_2, ..., z_n \mid c_i)}{P(z_1, z_2, ..., z_n)} \text{ (theo định lý Bayes)}$$

Để tìm được phân lớp có thể nhất đối với z thì:

$c_{MAP} = \arg\max_{c_i \in C} P(z_1, z_2, ..., z_n \mid c_i).P(c_i)$ với $P(z_1, z_2, ..., z_n)$ là như nhau với các lớp.

Giả sử (assumption): Các thuộc tính là độc lập có điều kiện (conditionally independent) đối với các lớp:

$$P(z_1, z_2, ..., z_n \mid c_i) = \prod_{j=1}^{n} P(z_j \mid c_i)$$

Thay vào công thức trên ta có Phân loại Naïve Bayes để tìm được phân lớp có thể nhất đối với z:

$$c_{NB} = \arg\max_{c_i \in C} P(c_i).\prod_{j=1}^{n} P(z_j \mid c_i)$$

Trong đó $P(z_j \mid c_i)$ được tính từ dữ liệu huấn luyện bằng số lần z_j xuất hiện cùng với c_i chia cho số lần z_j xuất hiện. Việc tính xác suất này đòi hỏi ít dữ liệu hơn nhiều so với tính $P(z_1, z_2, ..., z_n \mid c_i)$. Học Bayes đơn giản không đòi hỏi tìm kiếm trong không gian các bộ phân loại như đối với trường hợp học cây quyết định.

2.1.5. Giải thuật phân loại Naïve Bayes

a) Giai đoạn học (training phase), sử dụng một tập học

Đối với mỗi phân lớp có thể (mỗi nhãn lớp) $c_i \in C$

- Tính giá trị xác suất trước: $P(c_i)$
- Đối với mỗi giá trị thuộc tính x_j, tính giá trị xác suất xảy ra của giá trị thuộc tính đó đối với một phân lớp c_i: $P(x_j \mid c_i)$

b) Giai đoạn phân lớp (classification phase), đối với một ví dụ mới

Đối với mỗi phân lớp $c_i \in C$, tính giá trị của biểu thức: $P(c_i).\prod_{j=1}^{n} P(x_j \mid c_i)$

Xác định phân lớp của z là lớp có thể nhất c*

$$c^* = \arg\max_{c_i \in C} P(c_i).\prod_{j=1}^{n} P(x_j \mid c_i)$$

Ví dụ bài toán phân loại Naïve Bayes: Một sinh viên trẻ với mức thu nhập trung bình và mức đánh giá tín dụng bình thường có mua máy tính hay không?

Bảng 2.2. Số liệu dự đoán sinh viên mua máy tính

Rec ID	Age	Income	Student	Creadit Rating	Buy Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Medium	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Medium	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Excellent	Yes
12	Medium	Medium	No	Excellent	Yes
13	Medium	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

Biểu diễn bài toán phân loại

z = (Age=Young, Income=Medium, Student=Yes, Credit Rating=Fair)

Có 2 phân lớp có thể: c_1 (“Yes”) và c_2 (“No”)

Tính giá trị xác suất cho mỗi phân lớp $P(c_1)$ = 9/14 , $P(c_2)$ = 5/14

Tính giá trị xác suất của mỗi giá trị thuộc tính đối với mỗi phân lớp

- P(Age=Young| c_1) = 2/9; P(Age=Young| c_2) = 3/5
- P(Income=Medium| c_1) = 4/9; P(Income=Medium| c_2) = 2/5
- P(Student=Yes| c_1) = 6/9; P(Student=Yes| c_2) = 1/5
- P(Credit_Rating=Fair| c_1) = 6/9; P(Credit_Rating=Fair| c_2) = 2/5

Tính toán xác suất có thể xảy ra (likelihood) của ví dụ z đối với mỗi phân lớp

- Đối với phân lớp c_1

P(z| c_1) = P(Age=Young| c_1).P(Income=Medium| c_1).P(Student=Yes| c_1).

P(Credit_Rating=Fair| c_1) = (2/9).(4/9).(6/9).(6/9) = 0.044

- Đối với phân lớp c_2

P(z| c_2) = P(Age=Young| c_2).P(Income=Medium| c_2).P(Student=Yes| c_2).

P(Credit_Rating=Fair| c_2) = (3/5).(2/5).(1/5).(2/5) = 0.019

Xác định phân lớp có thể nhất (the most probable class)

- Đối với phân lớp c_1

P(c_1).P(z| c_1) = (9/14).(0.044) = 0.028

- Đối với phân lớp c_2

P(c_2).P(z| c_2) = (5/14).(0.019) = 0.007

→ Kết luận: **Anh ta (z) sẽ mua một máy tính!**

2.1.6. Các vấn đề trong phân loại Naïve Bayes và hướng xử lý

- Nếu không có ví dụ nào gắn với phân lớp c_i có giá trị thuộc tính x_j . $P(x_j \mid c_i)$ =0 , vì vậy $P(c_i).\prod_{j=1}^{n} P(x_j \mid c_i)$ =0

Giải pháp: Sử dụng phương pháp Bayes để ước lượng $P(x_j \mid c_i)$

$$P(x_j \mid c_i) = \frac{n(c_i, x_j) + mp}{n(c_i) + m}$$

$n(c_i)$: số lượng các ví dụ học gắn với phân lớp c_i

$n(c_i, x_j)$: số lượng các ví dụ học gắn với phân lớp c_i có giá trị thuộc tính x_j

p: xác suất tiền nghiệm của x_j, $p = 1/k$ trong đó *k* là số thuộc tính của thuộc tính X_i.

m: tham số cho phép xác định ảnh hưởng của p tới công thức.

- Giới hạn về độ chính xác trong tính toán của máy tính $P(x_j \mid c_i) < 1$, đối với mọi giá trị **t**huộc tính x_j và phân lớp c_i **. V**ì vậy, khi số lượng các giá trị thuộc tính là rất lớn, thì:

$$\lim_{n \to \infty} \prod_{j=1}^{n} P(x_j \mid c_i) = 0$$

Giải pháp: Sử dụng hàm lôgarit cho các giá trị xác suất.

$$c_{NB} = \arg\max_{c_i \in C} \left(\log \left[\mathrm{P}(\mathrm{c}_i).\prod_{j=1}^{n} P(\mathrm{x}_j \mid c_i) \right] \right)$$

$$c_{NB} = \arg\max_{c_i \in C} \left(\log\ \mathrm{P}(\mathrm{c}_i) + \sum_{j=1}^{n} \log P(\mathrm{x}_j \mid c_i) \right)$$

2.2. Thuật toán Support Vector Machine (SVM)[5][7]

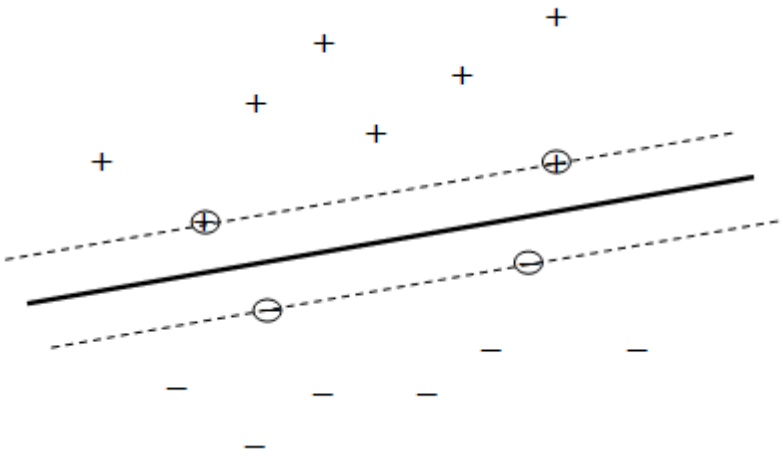
Support Vector Machines (SVM) là một phương pháp phân loại xuất phát từ lý thuyết học thống kê, dựa trên nguyên tắc tối thiểu rủi ro cấu trúc (Structural Risk Minimisation). SVM sẽ cố gắng tìm cách phân loại dữ liệu sao cho có lỗi xảy ra trên tập kiểm tra là nhỏ nhất (Test Error Minimisation). Đây là một phương pháp mới trong lĩnh vực trí tuệ nhân tạo. Vào thời kỳ đầu khi SVM xuất hiện, khả năng tính toán của máy tính còn rất hạn chế, nên phương pháp SVM không được lưu tâm. Tuy nhiên, từ năm 1995 trở lại đây, các thuật toán sử dụng cho SVM phát triển rất nhanh, cùng với khả năng tính toán mạnh mẽ của máy tính, đã có được những ứng dụng rất to lớn.

2.2.1. Ý tưởng

Cho trước một tập huấn luyện, được biểu diễn trong không gian vector, trong đó mỗi đối tượng là một điểm, phương pháp này tìm ra một siêu phẳng f quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp "+" và lớp "−". Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.

Ý tưởng của nó là ánh xạ (tuyến tính hoặc phi tuyến) dữ liệu vào không gian các vector đặc trưng (space of feature vectors) mà ở đó một siêu phẳng tối ưu được tìm ra để tách dữ liệu thuộc hai lớp khác nhau.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất:



Hình 2.1. Mô tả phương pháp SVM

Đường tô đậm là siêu phẳng tốt nhất và các điểm được bao bởi hình tròn là những điểm gần siêu phẳng nhất, chúng được gọi là các vector hỗ trợ (Support Vector). Các đường nét đứt mà các Support Vector nằm trên đó được gọi là lề (margin).

Các vector hỗ trợ (Support Vector) quyết định hàm phân tách dữ liệu. Từ đây, có thể thấy phương pháp SVM không phụ thuộc vào các mẫu dữ liệu ban đầu, mà chỉ phụ thuộc vào các suport vector (quyết định 2 siêu phẳng phân tách). Cho dù các điểm khác bị xoá thì thuật toán vẫn cho ra các kết quả tương tự. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác do các điểm trong tập dữ liệu đều được dùng để tối ưu kết quả.

2.2.2. Cơ sở lý thuyết

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian *F* và siêu phẳng quyết định f trên *F* sao cho sai số phân loại là thấp nhất.

Cho tập mẫu $D = \{(x_1, y_1), (x_1, y_1), ..., (x_i, y_i)\}$ với $x_i \in R^n$, thuộc vào hai lớp nhãn $y_i \in \{-1, 1\}$ là nhãn lớp tương ứng của các x_i (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vector $\vec{x_i}$ trong không gian:

$$x_i^{\rightarrow}.\vec{w} + b = 0$$

Đặt $f(x_l^{\rightarrow}) = sign(x_i^{\rightarrow}.\overrightarrow{w} + b) = \begin{cases} +1, x_i^{\rightarrow}.\overrightarrow{w} + b > 0 \\ -1, x_i^{\rightarrow}.\overrightarrow{w} + b < 0 \end{cases}$

Như vậy, $f(\vec{x}_i)$ biểu diễn sự phân lớp của $\vec{x}_i$ vào hai lớp như đã nêu.

Ta nói $y_i = +1$ nếu $\vec{x}_i$ thuộc lớp I(+) và $y_i = -1$ nếu $\vec{x}_i$ thuộc lớp (-). Khi đó, để có siêu phẳng f ta sẽ phải giải bài toán sau:

Tìm $\min \|\overrightarrow{w}\|$ với $\overrightarrow{w}$ thỏa mãn điều kiện sau:

$$y_i(sign(\vec{x}_i.\vec{w} + b)) \geq 1 \text{ với } \forall i \in \overline{1,n}$$

Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrage để biến đổi về dạng đẳng thức. Một đặc điểm thú vị của SVM là mặt phẳng quyết định chỉ phụ thuộc các Support Vector và nó có khoảng cách đến mặt phẳng quyết định là $1/\|\overrightarrow{w}\|$. Cho dù các đặc điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống như ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả.

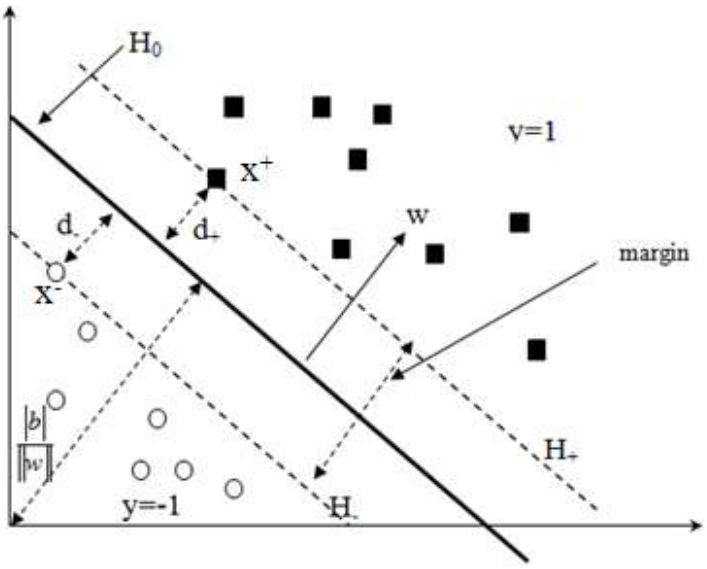
Tóm lại, trong trường hợp nhị phân phân tách tuyến tính, việc phân lớp được thực hiện qua hàm quyết định f(x) = sign(<w.x> + b), hàm này thu được bằng việc thay đổi vectơ chuẩn w, đây là vectơ để cực đại hóa viền chức năng $\gamma = \frac{1}{\|w\|_2}$.

2.2.3. Thuật toán SVM với bài toán phân hai lớp

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới x_i thì cần phải xác định x_i được phân vào lớp +1 hay lớp −1. Với bài toán phân hai lớp, ta có 3 trường hợp cần xét, mỗi trường hợp sẽ có 1 bài toán tối ưu, giải được bài toán tối ưu đó ta sẽ tìm được siêu phẳng cần tìm.

2.2.3.1. Thuật toán SVM trong trường hợp dữ liệu được phân tách tuyến tính

Tập D có thể phân chia tuyến tính được mà không có nhiễu (tất cả các điểm được gán nhãn +1 thuộc về phía dương của siêu phẳng, tất cả các điểm được gán nhãn -1 thuộc về phía âm của siêu phẳng).



Hình 2.2. Tập dữ liệu được phân chia tuyến tính

Xét hai điểm mẫu (x^+ ,+1) và (x^- ,-1) gần nhất đối với siêu phẳng phân tách $H_0\langle w . x \rangle + b=0$ như trong hình 2.7. Hai siêu phẳng lề song song với nhau:

$H_+\langle w . x \rangle + b =1$ đi qua x^+

$H_-\langle w . x \rangle + b =-1$ đi qua x^-

Cả hai siêu phẳng này đều song song với H_0 sao cho:

$$\begin{cases} \langle w.x_i \rangle + b \geq 0 \text{ nếu } y_i = 1 \\ \langle w.x_i \rangle + b \geq 0 \text{ nếu } y_i = -1 \end{cases}$$

Khi đó, mức lề (*margin*) là khoảng cách giữa hai siêu phẳng lề H_+ và H_- chính là khoảng cách $d_+ + d_-$. Trong đó, d_+ là khoảng cách giữa H_+ và H_0 và d_- là khoảng cách giữa H_- và H_0. Theo không gian vector trong đại số tuyến tính thì khoảng cách từ điểm x_i tới một siêu phẳng $\langle w . x \rangle + b =0$ là : $\dfrac{|\langle w. x_i \rangle + b|}{\|w\|}$

Trong đó $\|w\|$ là độ dài của vector w : $\|w\| = \sqrt{\langle w.w \rangle} = \sqrt{w_1^2 + w_2^2 + ... + w_r^2}$

Ta có : $d_+ = \dfrac{|\langle w.x^+ \rangle| + b}{w} = \dfrac{|1|}{\|w\|}$ và $d_- = \dfrac{|\langle w.x^- \rangle| + b}{w} = \dfrac{|-1|}{\|w\|} = \dfrac{|1|}{\|w\|}$

Khi đó mức lề (*margin*) : $d_+ + d_- = \dfrac{2}{\|w\|}$

Thuật toán SVM học một phân lớp nhằm cực đại hóa mức lề. Tương đương với giải quyết bài toán tối ưu bậc hai sau : Tìm w và b sao cho : $margin = \frac{2}{\|w\|}$ đạt cực đại.

Với mọi ví dụ huấn luyện x_i (i= 1,….,n) bài toán trên tương đương với bài toán cực tiểu hóa có ràng buộc sau :

Cực tiểu hóa : $\frac{\langle w.w \rangle}{2}$ với điều kiện : $y_i(\langle w.x_i \rangle + b) \geq 1, \ \forall i = 1,2,...n$.

Bài toán cực tiểu hóa có ràng buộc đẳng thức như sau :

Cực tiểu hóa $f(x)$, với điều kiện $g(x)$=0. Điều kiện cần để x_0 là một lời giải :

$$\begin{cases} \frac{\partial}{\partial x}\left(f(x) + \alpha\, g(x)\Big|_{x=x_0} = 0\right) \\ g(x) = 0 \end{cases}$$ Với α là một hệ số nhân Lagrange.

Trong trường hợp có nhiều ràng buộc đẳng thức $g_i(x)$=0 (i = 1…n), cần một hệ số nhân Lagrange cho mỗi ràng buộc :

$$\begin{cases} \frac{\partial}{\partial x}\left(f(x) + \sum_{i=1}^{r} \alpha_i\, g_i(x)\right)\Big|_{x=x_0} = 0 \\ g_i(x) = 0 \end{cases}$$ Với α_i là một hệ số nhân Lagrange.

Bài toán cực tiểu hóa có các ràng buộc bất đẳng thức : Cực tiểu hóa f(x), với các điều kiện $g_i(x) \leq 0$.

Điều kiện cần để x_0 là một lời giải :

$$\begin{cases} \frac{\partial}{\partial x}\left(f(x) + \sum_{i=1}^{r} \alpha_i\, g_i(x)\right)\Big|_{x=x_0} = 0 \\ g_i(x) = 0 \end{cases} \qquad \text{Với} : \alpha_i \geq 0 \qquad (2.1)$$

Hàm : $L = f(x) + \sum_{i=1}^{n} \alpha_i\, g_i(x)$ được gọi là hàm Lagrange đối với phương trình (2.1)

Biểu thức Lagrange được xây dựng bằng cách ràng buộc được nhân với các hệ số nhân dương và được trừ vào hàm mục tiêu, biểu thức Lagrange đối với bài toán là :

$$L_p(w,b,a) = \frac{1}{2}\langle w.w\rangle - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle w.x_i\rangle + b\right) - 1\right] \qquad (2.2)$$

Trong đó $\alpha_i \geq 0$ là các hệ số nhân Lagrange.

Lời giải tối ưu cho (2.2) phải thỏa mãn các điều kiện nhất định gọi là các điều kiện Karush –Kuhn-Tucker(KKT- Các điều kiện cần nhưng không phải là các điều kiện đủ). Các điều kiện KKT đóng vai trò trung tâm trong cả lý thuyết và ứng dụng của lĩnh vực tối ưu có ràng buộc.

Tập điều kiện KKT :

$$\frac{\partial L_p}{\partial w_j} = w_j - \sum_{i=1}^{n} y_i\alpha_i x_{ij} = 0 \ , j = 1,2,...,r$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n} y_i\alpha_i = 0$$

$$y_i\left(\langle w.x_i\rangle + b\right) - 1 \geq 0, \ i = 1,2,...,n$$

$$\alpha_i \geq 0, \ i = 1,2,...,n$$

$$\alpha_i\left[y_i\left(\langle w.x_i\rangle + b\right) - 1\right] = 0, \ i = 1,2,...,n$$

Phương pháp Lagrange giải quyết bài toán tối ưu hàm lồi dẫn đến một biểu thức đối ngẫu của bài toán tối ưu dễ giải quyết hơn so với biểu thức cần tối ưu ban đầu.

Cách thu được biểu thức đối ngẫu từ biểu thức ban đầu: gán giá trị bằng 0 đối với các đạo hàm bộ phận của biểu thức Lagrange (2.2) đối với các biến ban đầu (w và b), sau đó áp dụng các quan hệ thu được đối với biểu thức Lagrange ban đầu (2.2).

Biến đối ngẫu L_D :

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i.x_j\rangle$$

Cả hai biểu thức L_p và L_D đều là các biểu thức Lagrange dựa trên cùng một hàm mục tiêu nhưng với các ràng buộc khác nhau. Lời giải tìm được bằng cách cực tiểu hóa L_p hoặc cực đại hóa L_D

Cực đại hóa : $L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i . x_j \rangle$

Với điều kiện $\begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ \alpha_i \geq 0, \ i = 1,2,...,n \end{cases}$

Gọi SV là tập các vector hỗ trợ, SV là tập con của tập n các ví dụ huấn luyện ban đầu $\rightarrow \alpha_i > 0$ đối với các vector hỗ trợ x_i và $\alpha_i = 0$ đối với các vector không phải là vector hỗ trợ x_i .

Sử dụng các điều kiện Karush-Kuhn-Tucker ta tính được giá trị w và b, từ đó suy ra công thức siêu phẳng quyết định ranh giới phân lớp:

$$f(x) = \langle w.x \rangle + b = \sum_{x_i \in SV} \alpha_i y_i \langle x_i . x \rangle + b = 0$$

Đối với một ví dụ cần phân lớp z, cần tính giá trị :

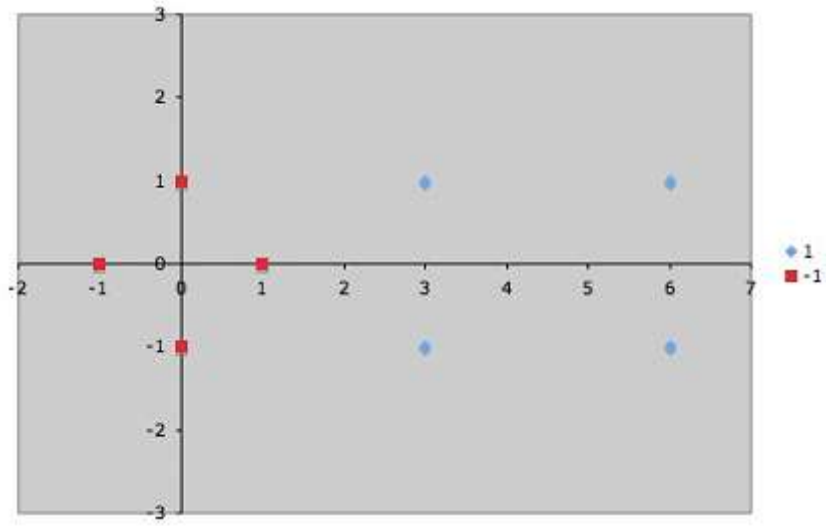
$$Sign(\langle w.x \rangle + b) = Sign\left(\sum_{x_i \in SV} \alpha_i y_i \langle x_i . x \rangle + b = 0 \right) \qquad (2.3)$$

Nếu biểu thức (2.3) trả về giá trị 1 thì ví dụ z được phân vào lớp có nhãn dương, ngược lại thì được phân vào lớp có nhãn âm. Việc phân lớp này chỉ phụ thuộc vào các vector hỗ trợ và chỉ cần giá trị tích vô hướng của hai vector chứ không cần biết giá trị hai vector đấy.

Ví dụ trường hợp dữ liệu phân tách tuyến tính: Giả sử có mặt phẳng R^+

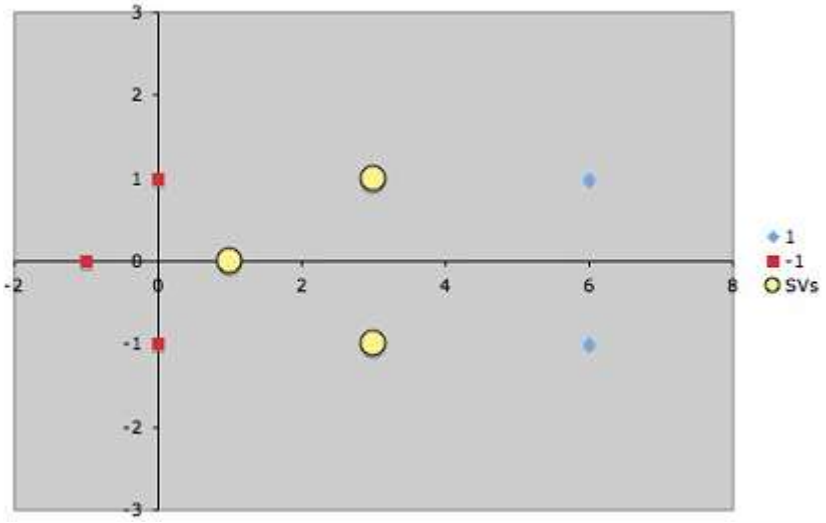
Tập các điểm được gán nhãn dương (+1): {(3,1), (3, -1), (6, 1), (6, -1)}

Và tập các điểm được gán nhãn âm (-1): {(1, 0), (0, 1), (0, -1), (-1, 0)}



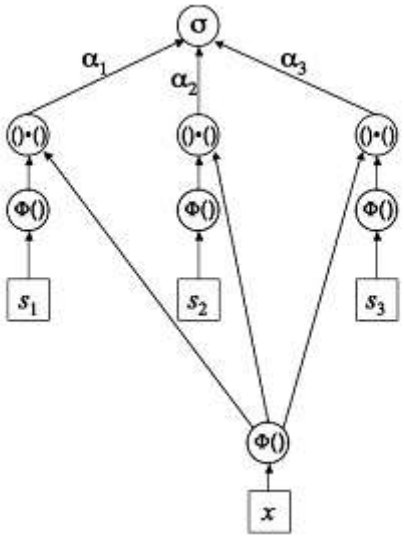
Hình 2.3. Các điểm dữ liệu được biểu diễn trên R^+

Dùng SVM để phân biệt hai lớp (+1 và -1). Bởi vì dữ liệu được chia tách một cách tuyến tính rõ ràng, nên sử dụng linear SVM (SVM tuyến tính) để thực hiện. Theo quan sát hình 2.4, chọn ra 3 vector hỗ trợ để thực thi các phép toán nhằm tìm ra mặt phẳng phân tách tối ưu nhất: $\{s_1 = (1,0), s_2 = (3,1), s_3 = (3, -1)\}$



Hình 2.4. : Các vector hỗ trợ (support vector) được chọn

Các vector hỗ trợ được tăng cường bằng cách thêm 1. Tức là $s_1 = (1, 0)$, thì nó sẽ được chuyển đổi thành $\tilde{s} = (1, 0, 1)$.



Hình 2.5. Mô hình kiến trúc SVM

Theo kiến trúc SVM, cần tìm ra những giá trị α_i.

$$\alpha_1\Phi(s_1).\Phi(s_1)+\alpha_2\Phi(s_2).\Phi(s_1)+\alpha_3\Phi(s_3).\Phi(s_1)=-1$$
$$\alpha_1\Phi(s_1).\Phi(s_2)+\alpha_2\Phi(s_2).\Phi(s_2)+\alpha_3\Phi(s_3).\Phi(s_2)=+1$$
$$\alpha_1\Phi(s_1).\Phi(s_1)+\alpha_2\Phi(s_2).\Phi(s_3)+\alpha_3\Phi(s_3).\Phi(s_3)=+1$$

Bởi vì sử dụng SVM tuyến tính nên hàm Φ dùng để chuyển đổi vector từ không gia dữ liệu đầu vào sang không gian đặc trưng . Biểu thức trên được viết lại như sau:

$$\alpha_1\tilde{s}_1.\tilde{s}_1+\alpha_2\tilde{s}_2.\tilde{s}_1+\alpha_3\tilde{s}_3.\tilde{s}_1=-1$$
$$\alpha_1\tilde{s}_1.\tilde{s}_2+\alpha_2\tilde{s}_2.\tilde{s}_2+\alpha_3\tilde{s}_3.\tilde{s}_2=+1$$
$$\alpha_1\tilde{s}_1.\tilde{s}_3+\alpha_2\tilde{s}_2.\tilde{s}_3+\alpha_3\tilde{s}_3.\tilde{s}_3=+1$$

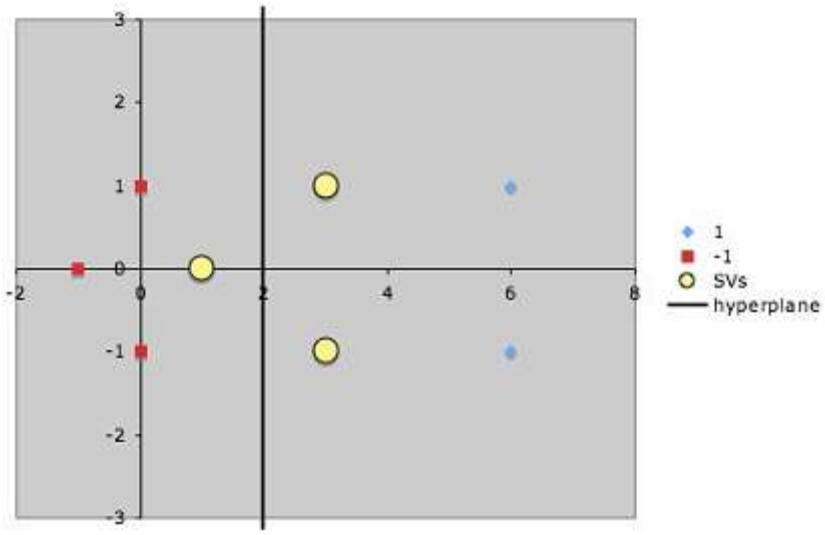
Rút gọn biểu thức trên thông qua việc tính toán tích vô hướng giữa các vector.

$$2\alpha_1+4\alpha_2+4\alpha_3=-1$$
$$4\alpha_1+11\alpha_2+9\alpha_3=+1$$
$$4\alpha_1+9\alpha_2+11\alpha_3=+1$$

Giải hệ phương trình 3 ẩn trên : $\alpha_1 = -3.5$, $\alpha_2 = 0.75$, $\alpha_3 = 0.75$. Tiếp đến tính trọng số $\tilde{\omega}$ thông qua công thức:

$$\tilde{\omega} = \sum_i \alpha_i \tilde{s}_i = -3.5\begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\-1\\1\end{pmatrix} = \begin{pmatrix}1\\0\\-2\end{pmatrix}$$

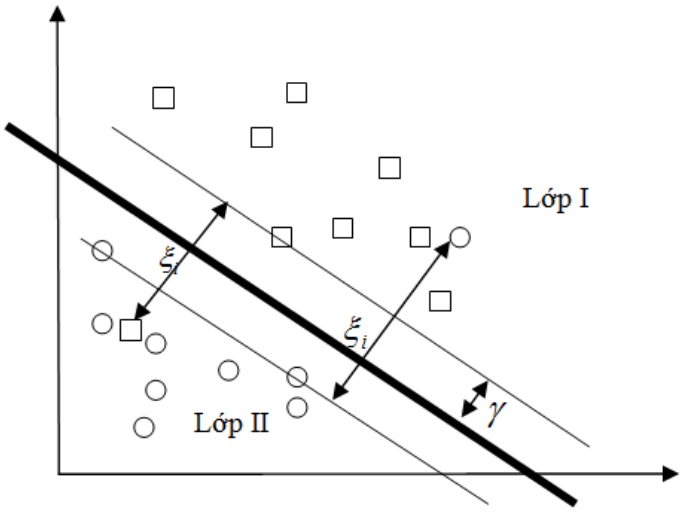
Siêu phẳng phân chia 2 lớp đó là: y = wx + b với w = (1, 0) và b = -2



Hình 2.6. Siêu phẳng được biểu diễn trên R^+

2.2.3.2. Thuật toán SVM trong trường hợp dữ liệu có nhiễu

Tập dữ liệu D có thể phân chia tuyến tính được nhưng có nhiễu. Trong trường hợp này, hầu hết các điểm đều được phân chia đúng bởi siêu phẳng. Tuy nhiên có 1 số điểm bị nhiễu, nghĩa là: Điểm có nhãn dương nhưng lại thuộc phía âm của siêu phẳng, điểm có nhãn âm nhưng lại thuộc phía dương của siêu phẳng.



Hình 2.7. Tập dữ liệu phân chia tuyến tính nhưng có nhiễu

Để làm việc với các dữ liệu nhiễu, cần nới lỏng các điều kiện lề bằng cách sử dụng các biến slack $\xi_i (\geq 0)$ như sau:

$$\langle w.x_i \rangle + b \geq 1 - \xi_i \quad \text{nếu } y_i = 1$$

$$\langle w.x_i \rangle + b \leq -1 + \xi_i \quad \text{nếu } y_i = -1$$

Đối với một ví dụ nhiễu/lỗi: $\xi_i > 1$ thì $\sum_i \xi_i$ là giới hạn trên của lỗi của các ví dụ huấn luyện. Các điều kiện đối với trường hợp dữ liệu tuyến tính không phân tách được :

$$\begin{cases} y_i \left(\langle w.x_i \rangle + b \right) \geq 1 - \xi_i \ , i = 1,2,...,n \\ \xi_i \geq 0 \ , i = 1,2,...,n \end{cases}$$

Cần phải tích hợp lỗi trong hàm tối ưu mục tiêu, bằng cách gán giá trị chi phí (cost) cho các lỗi, và tích hợp chi phí này trong hàm mục tiêu mới.

Chúng ta cần cực tiểu hóa:

$$\frac{\langle w.w \rangle}{2} + C \left(\sum_{i=1}^{n} \xi_i \right)^k$$

Với $C \geq 0$ là tham số xác định mức độ chi phí. K = 1 được sử dụng phổ biến, nó có tiến bộ là không chứa ξ_i và các hệ số nhân Lagrange, chúng ta chỉ quan tâm trường hợp k = 1.Bài toán tối ưu hóa mới trở thành:

Cực tiểu hóa : $\frac{\langle w.w \rangle}{2} + C \sum_{i=1}^{n} \xi_i$

Với điều kiện :

$$\begin{cases} y_i \left(\langle w.x_i \rangle + b \right) \geq 1 - \xi_i \ , i = 1,2,...,n \\ \xi_i \geq 0 \ , i = 1,2,...,n \end{cases}$$

Trong đó C(> 0) là tham số xác định mức độ chi phí đối với các lỗi. Giá trị C càng lớn thì mức độ chi phí lỗi càng cao. Bài toán tối ưu mới này được gọi là Soft-margin SVM.

Biểu thức tối ưu Lagrange :

$$L_p = \frac{1}{2} \langle w.w \rangle + C \left(\sum_{i=1}^{n} \xi_i \right) - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\langle w.x_i \rangle + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{n} \mu_i \xi_i \qquad (2.4)$$

Trong đó, $\alpha_i, \mu_i \geq 0$ là các hệ số nhân Lagrange. Điều kiện KKT là những điều kiện sau :

$$\frac{\partial L_p}{\partial w_j} = w_j - \sum_{i=1}^{n} y_i \alpha_i x_{ij} = 0 \ , j = 1,2,...,r$$

(2.5)

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n} y_i \alpha_i = 0 \tag{2.6}$$

$$\frac{\partial L_p}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \ , i = 1,2,...,n \tag{2.7}$$

$$y_i \left(\langle w.x_i \rangle + b \right) - 1 + \xi_i \geq 0, \ i = 1,2,...,n \tag{2.8}$$

$$\alpha_i \geq 0, \ i = 1,2,...,n \tag{2.9}$$

$$\xi_i \geq 0, \ i = 1,2,...,n \tag{2.10}$$

$$\mu_i \geq 0, \ i = 1,2,...,n \tag{2.11}$$

$$\alpha_i \left[y_i \left(\langle w.x_i \rangle + b \right) - 1 + \xi_i \right] = 0, \ i = 1,2,...,n \tag{2.12}$$

$$\mu_i \xi_i = 0 \ , i = 1,2,...,n \tag{2.13}$$

Như trong trường hợp tuyến tính có thể phân tách, chuyển biểu thức Lagrange từ dạng ban đầu về dạng đối ngẫu bằng cách gán giá trị 0 cho các đạo hàm bộ phận của biểu thức Lagrange (2.4) đối với các biến ban đầu w, b, ξ_i . Sau đó thay thế các kết quả thu được vào biểu thức Lagrange ban đầu (2.4). Ta thu được biểu thức đối ngẫu $L_D(\alpha)$.

Bài toán trở thành :

Cực đại hóa : $$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i . x_j \rangle \tag{2.14}$$

Với điều kiện : $$\sum_{i=1}^{n} y_i \alpha_i = 0 \tag{2.15}$$
$$0 \leq \alpha_i \leq C, \ \forall i = 1,2,...,n$$

Bài toán đối ngẫu (2.14) cũng có thể được giải quyết bằng phương pháp số học, và kết quả các giá trị α_i sau đó được sử dụng để tính *w* và *b*.*w* được tính bởi phương trình (2.5) và *b* được tính bởi điều kiện bổ sung KKT (2.12) và (2.13). Từ phương trình (2.7), (2.12), (2.13) nhận thấy nếu $0 \le \alpha_i \le C$ thì cả $\xi_i = 0$ và $y_i\left(\langle w.x_i\rangle + b\right) - 1 + \xi_i = 0$. Vì thế có thể sử dụng bất kỳ điểm dữ liệu huấn luyện nào cho $0 \le \alpha_i \le C$ và phương trình (2.12) (với $\xi_i = 0$) để tính b :

$$b = \frac{1}{y_i} - \sum_{i=1}^{n} y_i \alpha_i \langle x_i . x_j \rangle \qquad (2.16)$$

Ta có thể tính tất cả các trường hợp có thể có của b và sau đó lấy giá trị trung bình để có được giá trị b cuối cùng.

Siêu phẳng quyết định phân tách :

$$f(x) = \langle w.x\rangle + b = \sum_{i=1}^{n} \alpha_i y_i \langle x_i . x\rangle + b$$

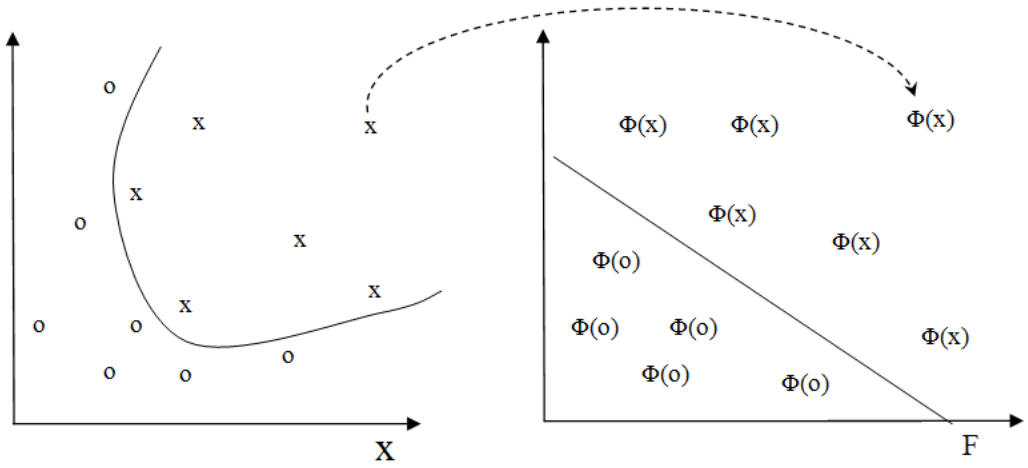
Luật quyết định cho phân lớp giống như trường hợp có thể phân tách được. Đối với một ví dụ cần phân lớp z, cần tính giá trị: $Sign(\langle w.x\rangle + b)$

2.2.3.3. Thuật toán SVM trong trường hợp dữ liệu phân tách phi tuyến

Tập dữ liệu D không thể phân chia tuyến tính được, ta sẽ ánh xạ các vector dữ liệu *x* từ không gian *n* chiều vào một không gian *m* chiều (*m*>*n*), sao cho trong không gian *m* chiều, D có thể phân chia tuyến tính được.

Không gian đặc trưng ký hiệu là F:

$$F = \{\Phi(x) \mid x \in X\}$$



Hình 2.8. Ánh xạ Φ từ không gian dữ liệu X sang không gian đặc trưng F

Hình 2.8 cho ta tập dữ liệu ban đầu không thể phân tách tuyến tính, ta sẽ xử lý bằng cách ánh xạ tập dữ liệu đã cho vào một không gian mới có số chiều lớn hơn không gian cũ (Gọi là không gian đặc trưng) mà trong không gian này tập dữ liệu có thể phân tách tuyến tính. Trong không gian đặc trưng ta sẽ tiếp tục tìm siêu phẳng phân tách.

Ý tưởng cơ bản là việc ánh xạ không gian biểu diễn: biểu diễn dữ liệu từ không gian ban đầu X sang một không gian đặc trưng F bằng cách áp dụng một hàm ánh xạ phi tuyến Φ.

$$\Phi : X \to F$$
$$x \mapsto \Phi(x)$$

Trong không gian đã chuyển đổi, tập các ví dụ học ban đầu

$$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

Được ánh xạ tương ứng thành:

$$\{(\Phi(x_1), y_1), (\Phi(x_2), y_2), ..., (\Phi(x_n), y_n)\}$$

Sau quá trình chuyển đổi không gian biểu diễn, bài toán tối ưu là.

Cực tiểu hóa: $L_p = \frac{\langle w.w \rangle}{2} + C\sum_{i=1}^{n} \xi_i$

Với điều kiện: $\begin{cases} y_i \left(\langle w.\Phi(x_i) \rangle + b \right) \geq 1 - \xi_i \ , i = 1,2,...,n \\ \xi_i \geq 0 \ , i = 1,2,...,n \end{cases}$

Bài toán tối ưu (đối ngẫu) tương ứng sẽ là:

Cực đại hóa: $L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \Phi(x_i).\Phi(x_j) \rangle$ (2.17)

Với điều kiện: $\begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ 0 \le \alpha_i \le C, \ \forall i = 1,2,...,n \end{cases}$

Ranh giới quyết định phân lớp là siêu phẳng phân tách

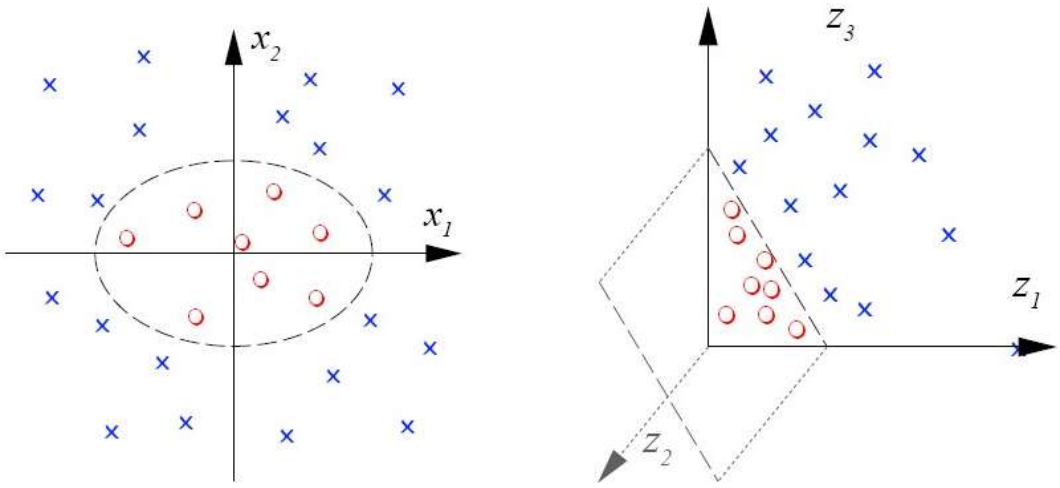
$$f(x) = \langle w.\Phi(x) \rangle + b = \sum_{i=1}^{n} \alpha_i y_i \langle \Phi(x_i).\Phi(x) \rangle + b \quad (2.18)$$

Trong biểu thức đối ngẫu (2.17) và trong biểu thức hàm phân tách tối ưu (2.18) việc xác định cụ thể giá trị $\Phi(x)$ và $\Phi(z)$ là không cần thiết mà chỉ cần tính giá trị tích vô hướng vector $\langle \Phi(x) . \Phi(z) \rangle$. Nếu có thể tính được tích vô hướng $\langle \Phi(x) . \Phi(z) \rangle$ trực tiếp từ các vector x và z thì không cần phải xác định vector đặc trưng trong không gian sau khi ánh xạ $\Phi(x)$ và hàm ánh xạ Φ. Trong phương pháp SVM, mục tiêu này đạt được thông qua việc sử dụng các hàm nhân (kernel functions), được ký hiệu là K.

$$K(x,z) = \langle \Phi(x).\Phi(z) \rangle$$

Ví dụ: Xét phép biến đổi dữ liệu từ không gian đầu vào $X = R^2$ vào không gian đặc trưng $F = \mathrm{R}^3$ được cho bởi:

$$\Phi : \mathrm{R}^2 \to \mathrm{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$$



Hinh 2.9. Ví dụ hàm hạt nhân

Khi ta chọn x = (x1, x2) và z = (z1, z2) ta có:

$$\langle x.z\rangle^2 = (\sum_{i=1}^{2} x_i, z_i)^2 = (z_1 z_1 + x_2 z_2)^2$$

$$= x_1^{\,2} z_1^{\,2} + 2x_1 z_1 x_2 z_2 + x_2^{\,2} z_2^{\,2}$$

$$= \left\langle (x_1^2, \sqrt{2}x_1 x_2, x_2^2).(z_1^2, \sqrt{2}z_1 z_2, z_2^2) \right\rangle$$

$$= \langle \Phi(x).\Phi(z) \rangle$$

Ta thấy khi chọn $K(x,z) = \langle \Phi(x).\Phi(z) \rangle = \langle x.z \rangle^2$. Vì vậy ta không cần phải tính ánh xạ Φ.

Theo 2.18, vì không cần xây dựng tường minh ánh xạ Φ và sử dụng hàm hạt nhân nên:

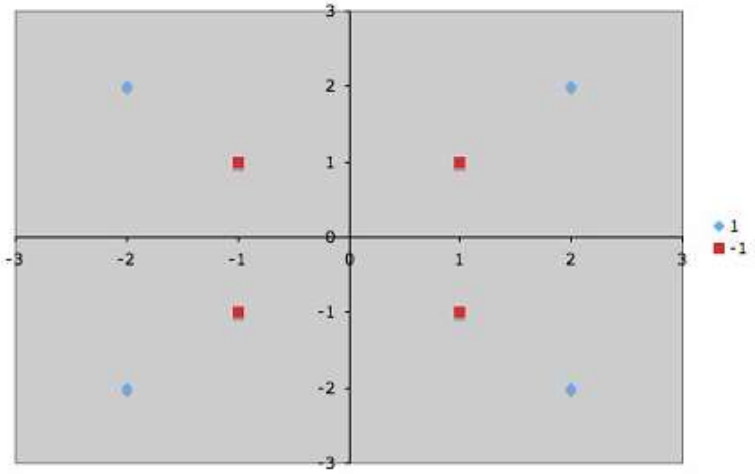
$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b$$

Một số hàm hạt nhân thường được sử dụng:

- $K(x,z) = \langle x.z \rangle$.
- $K(x,z) = (1 + \langle x.z \rangle)_{đ}^{d}$ với d là tham số do người dùng định nghĩa.

- $K(x,z)=\frac{1}{1-\langle x.z\rangle}$, $-1<\langle x.z\rangle<1$. với d là tham số do người dùng định nghĩa
- $K(x,z)=\exp(-y|x-z|^2)$, với y do người dùng định nghĩa

Ví dụ trường hợp dữ liệu phân tách phi tuyến: Giả sử ta có một tập các điểm được gán nhãn dương (+1): {(2,2), (2, -2), (-2, -2), (-2, 2)}. Và tập các điểm được gán nhãn âm (-1) trong mặt phẳng R^+: {(1, 1), (1, -1), (-1, -1), (-1, 1)}



Hình 2.10. Các điểm không phân chia tuyến tính

Có thể thấy rằng, trong không gian ban đầu, các điểm không được phân tách tuyến tính. Mục tiêu đặt ra là tìm một siêu phẳng có thể phân chia các điểm thành hai phần. Vì vậy phải sử dụng hàm ánh xạ $\Phi_1\left(x_1\ x_2\right)$ là một ánh xạ không tuyến tính từ không gian ban đầu sang một không gian đặc trưng.

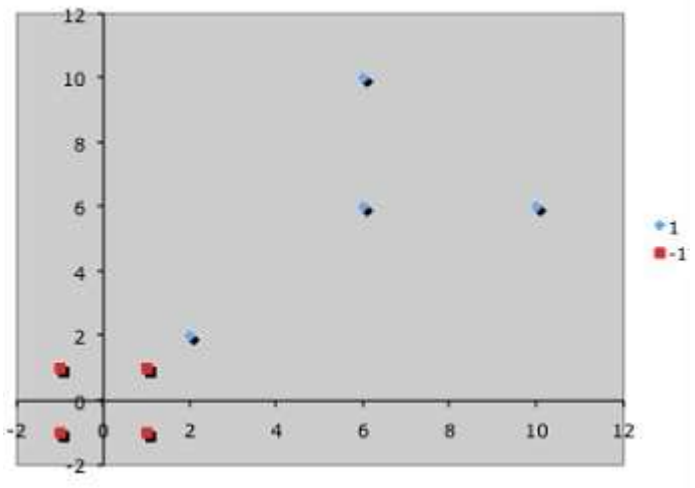
Định nghĩa

$$\Phi_1\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{cases}\begin{pmatrix}4-x_2+|x_1-x_2|\\4-x_1+|x_1-x_2|\end{pmatrix} & ,\sqrt{x_1^2+x_2^2}>2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix}\end{cases}$$

Theo kiến trúc SVM ta có thể biểu diễn lại dữ liệu trên không gian đặc trưng như sau:

Các điểm gán nhãn (+1) { (2,2) , (2,6) , (6,6) , (2,6) }

Các điểm gán nhãn (-1) { (1,1) , (1,-1) , (-1,-1) , (-1,1) }



Hình 2.11. Dữ liệu được biểu diễn lại trong không gian đặc trưng

Theo quan sát hình 2.11 , ta chọn ra 2 vector hỗ trợ để thực thi các phép toán nhằm tìm ra mặt phẳng phân tách tối ưu nhất: {s_1= (1,1), s_2 = (2,2)}

Các vector hỗ trợ được tăng cường bằng cách thêm 1. Tức là s_1 = (1, 1), thì nó sẽ được chuyển đổi thành $\tilde{s}$= (1, 1, 1). Theo kiến trúc SVM, công việc của chúng ta là tìm ra những giá trị α_i.

$$\alpha_1\Phi_1(s_1).\Phi_1(s_1) + \alpha_2\Phi_1(s_2).\Phi_1(s_1) = -1$$
$$\alpha_1\Phi_1(s_1).\Phi_1(s_2) + \alpha_2\Phi_1(s_2).\Phi_1(s_2) = +1$$

Trên không gian đặc trưng, biểu thức được viết lại như sau:

$$\alpha_1\tilde{s}_1.\tilde{s}_1 + \alpha_2\tilde{s}_2.\tilde{s}_1 = -1$$
$$\alpha_1\tilde{s}_1.\tilde{s}_2 + \alpha_2\tilde{s}_2.\tilde{s}_2 = +1$$

Rút gọn biểu thức trên thông qua việc tính toán tích vô hướng giữa các vector:

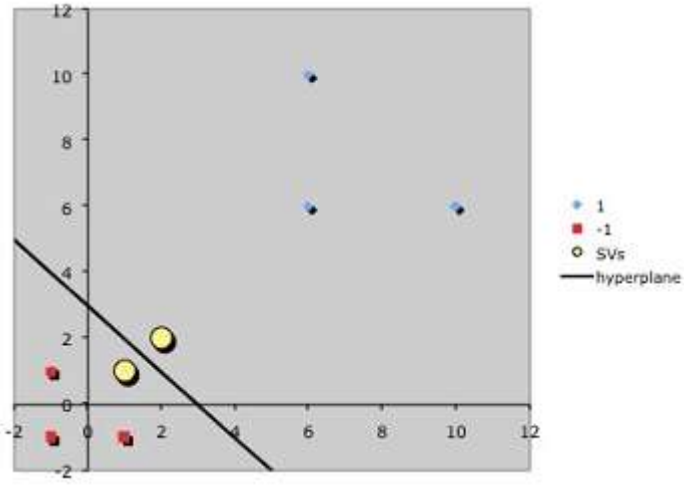
$$3\alpha_1 + 5\alpha_2 = -1$$
$$5\alpha_1 + 9\alpha_2 = +1$$

Giải hệ phương trình trên ta có α_1 = -7, α_2 = 4

Tiếp đến ta tính trọng số $\tilde{\omega}$ thông qua công thức:

$$\tilde{\omega} = \sum_i \alpha_i \tilde{s}_i = -7\begin{pmatrix}1\\1\\1\end{pmatrix} + 4\begin{pmatrix}2\\2\\1\end{pmatrix} = \begin{pmatrix}1\\1\\-3\end{pmatrix}$$

Siêu phẳng phân chia 2 lớp đó là: y = wx + b với w = (1, 1) và b = -3



Hình 2.12. Siêu phẳng phân tách tương ứng với giá trị α_1 = -7, α_2 = 4

Giả sử ta muốn phân loại điểm x= (4,5) sử dụng hàm ánh xạ $\Phi_1\left(x_1\ x_2\right)$. Theo công thức 2.18, phân lớp f(x) bằng cách giải phương trình sau:

$$f(x) = \sigma\sum_i \alpha_i \Phi(s_i).\Phi(x)$$

ở đây $\sigma(z)$ chính là dấu của z.

$$f\begin{pmatrix}4\\5\end{pmatrix} = \sigma\left(-7\Phi_1\begin{pmatrix}1\\1\end{pmatrix}.\Phi_1\begin{pmatrix}4\\5\end{pmatrix} + 4\Phi_1\begin{pmatrix}2\\2\end{pmatrix}.\Phi_1\begin{pmatrix}4\\5\end{pmatrix}\right)$$

$$= \sigma\left(-7\begin{pmatrix}1\\1\\1\end{pmatrix}.\begin{pmatrix}0\\1\\1\end{pmatrix} + 4\begin{pmatrix}2\\2\\1\end{pmatrix}.\begin{pmatrix}0\\1\\1\end{pmatrix}\right)$$

$$= \sigma(-2)$$

Vậy điểm x sẽ được phân vào lớp (-1).

2.2.4. Thuật toán SVM với bài toán phân đa lớp [8]

Để phân đa lớp thì kỹ thuật SVM sẽ chia không gian dữ liệu thành 2 phần và tiếp tục với không gian đã được phân chia. Khi đó hàm quyết định phân dữ liệu vào lớp thứ i sẽ là:

$$f_i(x) = \mathrm{w}_i^T x + b_i$$

Những phần tử x là support vector nếu thoả điều kiện:

$$f_i(x) = \begin{cases} 1, \in i \\ -1 \notin i \end{cases}$$

Giả sử bài toán phân loại *k* lớp ($k \geq 2$), ta sẽ tiến hành $k(k-1)/2$ lần phân lớp nhị phân sử dụng phương pháp SVM. Mỗi lớp sẽ tiến hành phân tách với *k-1* lớp còn lại để xác định *k-1* hàm phân tách (chiến lược "một-đối-một" (one-against-one).

Kỹ thuật phân đa lớp bằng phương pháp SVM hiện vẫn đang được tiếp tục nghiên cứu và phát triển.

2.2.5. Các bước chính của phương pháp SVM

- Tiền xử lý dữ liệu: Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thực thì ta cần phải tìm cách chuyển chúng về dạng số của SVM. Tránh các số quá lớn, thường nên co giãn dữ liệu để chuyển về đoạn [-1,1] hoặc [0,1].
- Chọn hàm hạt nhân: Cần chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.
- Thực hiện việc kiểm tra để xác định các tham số cho ứng dụng.
- Sử dụng các tham số cho việc huấn luyện tập mẫu.
- Kiểm thử tập dữ liệu Test.

2.2.6. So sánh và một số cải tiến

Một số phương pháp như mạng neuron, fuzzy logic, mạng fuzzy-neuron, …, cũng được sử dụng thành công để giải quyết bài toán phân lớp. Ưu điểm của các phương pháp này là không cần xác định mô hình toán đối của đối tượng (Giải quyết tốt với các hệ thống lớn và phức tạp).

SVM có 2 đặc trưng cơ bản:

- Luôn kết hợp với các dữ liệu có ý nghĩa về mặt vật lý, do vậy dễ dàng giải thích được một cách tường minh.
- Cần một tập các mẫu huấn luyện rất nhỏ.

Phương pháp SVM hiện nay được xem là một công cụ mạnh và tinh vi nhất hiện nay cho những bài toán phân lớp phi tuyến. Nó có một số biến thể như C – SVC, v – SVC. Cải tiến mới nhất hiện nay của phương pháp SVM đã được công bố là thuật toán NNSRM (Nearest Neighbor and Structural Risk Minimization) là sự kết hợp giữa 2 kỹ thuật SVM và Nearest Neighbor.

2.3. Xây dựng mô hình phân loại khách hàng dựa trên Naïve Bayes và Support Vector Machine (SVM)

2.3.1. Bài toán phân loại khách hàng dựa trên học máy

Phân tích yêu cầu

Luận văn mang ý nghĩa nhằm tạo ra sự khác biệt trong công tác chăm sóc khách hàng tại VNPT Hà Nội. Việc nghiên cứu và triển khai thành công sẽ đem lại ý nghĩa thiết thực, giúp nhà cung cấp VPNT Hà Nội trong hoạch định chiến lược phát triển. Mục tiêu của bài toán là phân loại khách hàng và dự đoán hành vi của khách hàng khôi phục hoặc tháo hủy dịch vụ để VNPT Hà Nội có phương án hỗ trợ, phát triển và giữ khách hàng. Vì các hành vi, yêu cầu của khách hàng được biểu hiện chính trên trạng thái của thuê bao khách hàng sử dụng nên luận văn cần giải quyết bài toán phân loại trạng thái của dịch vụ, dự đoán khả năng tháo hủy thuê bao của khách hàng.

Xác định yêu cầu

Hệ thống thực hiện được các chức năng:

- Xây dựng được mô hình phân loại các đối tượng khách hàng, thuê bao có nguy cơ thực hiện tháo hủy để xây dựng các chính sách ưu đãi mới cho các nhóm đối tượng này.
- Dự đoán khả năng tạm dừng dịch vụ của từng khách hàng, thuê bao để có chính sách riêng cho từng trường hợp.

- Đánh giá được mức độ ảnh hưởng của từng thuộc tính tới hành vi của thuê bao (khách hàng).
- Tỷ lệ lỗi dự đoán phải nằm trong mức cho phép . Theo yêu cầu tỷ lệ lỗi phải dưới 5% (Tỷ lệ dự đoán chính xác phải lớn hơn 95%).
- Hệ thống phải được xây dựng dựa trên cơ sở dữ liệu hiện có trên các hệ thống thông tin hiện có của VNPT Hà Nội, Dữ liệu phải tổng hợp tự động.
- Đáp ứng kịp thời và chính xác nhu cầu lấy số liệu của người dùng vào các thời điểm do người dùng quyết định.

Phạm vi bài toán

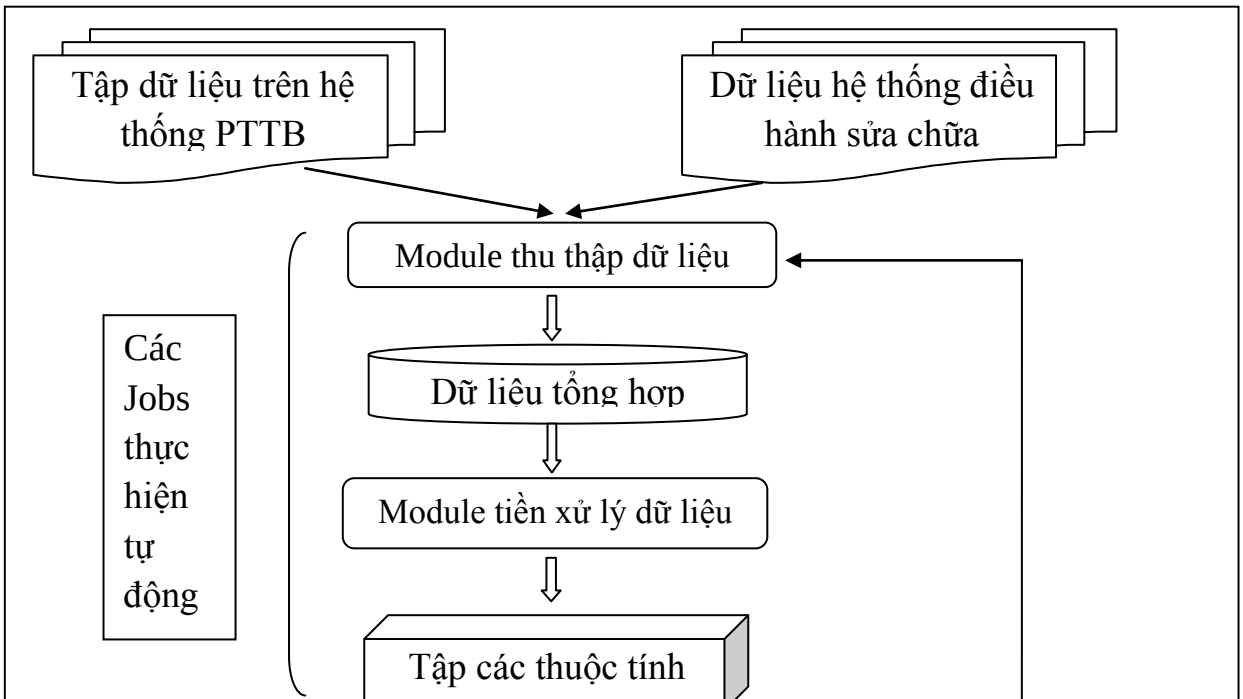
Dữ liệu bài toán là dữ liệu thông tin khách hàng và thông tin hợp đồng của VNPT Hà Nội từ khi triển khai hệ thống Điều hành phát triển thuê bao trên toàn địa bàn thành phố Hà Nội (Từ năm 2008 đến nay).

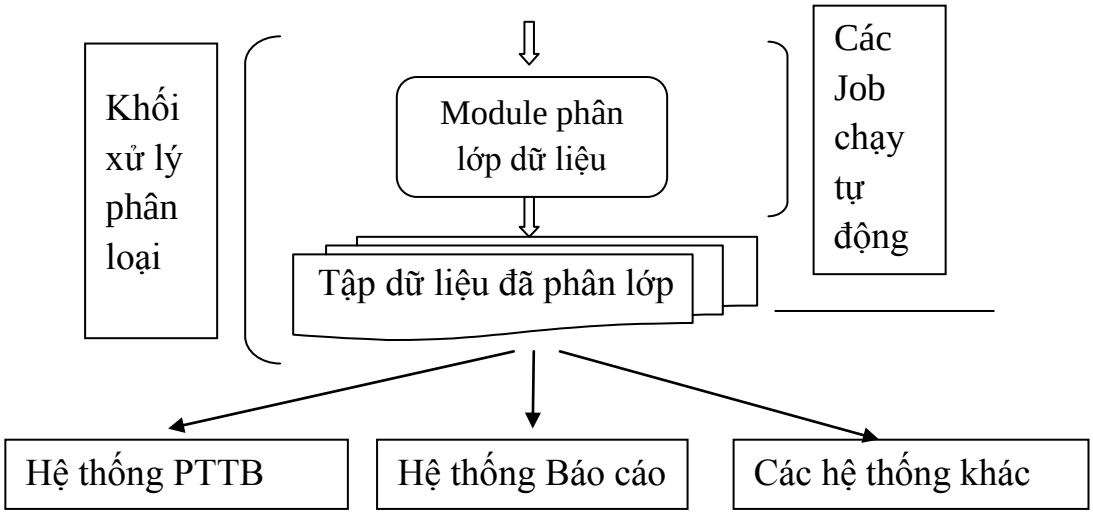
Phát biểu bài toán

Input: Tập các dữ liệu trên hệ thống Phát triển thuê bao của VNPT Hà Nội và các hệ thống khác như: Dữ liệu về thông tin khách hàng, dữ liệu về thông tin hợp đồng, thông tin về yêu cầu sửa chữa…

Output: Xác định khách hàng đó thuộc phân nhóm nào (có nguy cơ tạm dừng, tháo hủy hợp đồng hay không)

Để đáp ứng các yêu cầu bài toán phân loại đặt ra và áp dụng tại VNPT Hà Nội với dữ liệu có được từ hệ thống điều hành Phát triển thuê bao , luận văn đề xuất mô hình của toàn bộ hệ thống phân loại như hình 2.13.





Hình 2.13 : Mô hình phân loại khách hàng

2.3.2. Các bước xây dựng hệ thống

a) Thu thập dữ liệu

Thu thập dữ liệu phục vụ cho công việc phân loại là một khâu rất quan trọng, vì vậy cần một tập dữ liệu huấn luyện đủ lớn để áp dụng thuật toán học phân loại.

Tiến hành khảo sát hệ thống phát triển thuê bao để thu thập được những dữ liệu cho bài toán như:

- Dữ liệu về thông tin khách hàng.
- Dữ liệu về thông tin thuê bao .
- Dữ liệu về thông sửa chữa .
- Dữ liệu về hợp đồng thuê bao...

b) Tiền xử lý dữ liệu

Dữ liệu được chọn lọc sẽ phải qua bước tiền xử lý trước khi tiến hành khai phá phát hiện tri thức. Bước thu thập và tiền xử lý dữ liệu là bước rất phức tạp. Để một giải thuật khai phá dữ liệu thực hiện trên toàn bộ CSDL sẽ rất cồng kềnh, kém

hiệu quả. Trong quá trình khai phá dữ liệu, nhiều khi phải thực hiện liên kết/tích hợp dữ liệu từ rất nhiều nguồn khác nhau. Các hệ thống sẵn có được thiết kế với những mục đích và đối tượng phục vụ khác nhau, khi tập hợp dữ liệu từ những hệ thống này để phục vụ khai phá dữ liệu, hiện tượng dư thừa là rất phổ biến, ngoài ra còn có thể xảy ra xung đột gây mất dữ liệu, dữ liệu không đồng nhất, không chính xác. Rõ ràng yêu cầu chọn lọc và làm sạch dữ liệu là rất cần thiết.

Nếu đầu vào của quá trình khai phá là dữ liệu trong DataWarehouse (DW) thì sẽ rất thuận tiện, vì dữ liệu này đã được làm sạch, nhất quán và có tính chất hướng chủ đề.

Tuy nhiên nhiều khi vẫn phải có thêm một số bước tiền xử lý để đưa dữ liệu về đúng dạng cần thiết.

c) Phân lớp dữ liệu

Nhiệm vụ phân lớp bắt đầu với việc xây dựng dữ liệu (dữ liệu huấn luyện) có các giá trị đích (nhãn lớp) đã biết. Các thuật toán phân lớp khác nhau dùng các kỹ thuật khác nhau cho việc tìm các quan hệ giữa các giá trị của thuộc tính dự báo và các giá trị của thuộc tính đích trong dữ liệu huấn luyện. Những quan hệ này được tổng kết trong mô hình, sau đó được dùng cho các trường hợp mới với các giá trị đích chưa biết để dự đoán các giá trị đích.

Phân lớp sẽ được thực hiện theo 2 bước như sau

Bước 1. Xây dựng mô hình (Học)

Xây dựng mô hình bằng cách phân tích tập dữ liệu huấn luyện, sử dụng các thuật toán phân lớp và thể hiện mô hình theo thuật toán Naïve Bayes hoặc SVM

Bước này còn được coi là bước tạo ra bộ phân lớp (classifier).

Bước 2. Sử dụng mô hình (Phân lớp)

Áp dụng mô hình cho tập dữ liệu kiểm thử với các lớp đã xác định để kiểm tra và đánh giá độ chính xác của mô hình. Nếu độ chính xác là chấp nhận được, mô hình sẽ được sử dụng để phân lớp cho các dữ liệu mới.

Như vậy có 3 tập dữ liệu có cấu trúc và các thuộc tính dự đoán giống nhau: Tập huấn luyện , tập kiểm thử đã biết lớp và tập mới chưa xác định lớp.

d) Đánh giá hiệu quả của phương pháp.

2.4. Kết chương

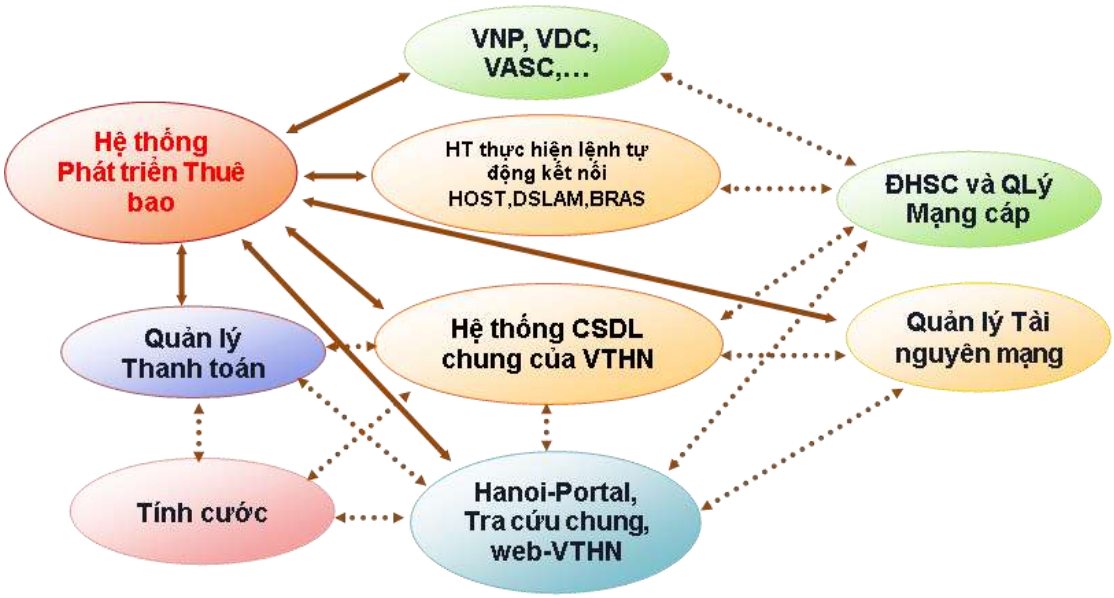
Chương 2 tập trung nghiên cứu về 2 thuật toán là Naïve Bayes và SVM để hiểu rõ việc thực hiện huấn luyện và phân loại đồng thời cũng mô tả mô hình phân loại khách hàng bằng hai thuật toán dựa trên dữ liệu thực tế về khách hàng và thuê bao của VNPT Hà Nội.

CHƯƠNG III: THỬ NGHIỆM VÀ KẾT QUẢ

3.1. Giới thiệu bộ dữ liệu thử nghiệm

3.1.1. Khái quát về hệ thống phát triển thuê bao của VNPT Hà Nội

Áp dụng công nghệ tin học vào công tác quản lý điều hành sản xuất kinh doanh từ rất lâu, đến nay VNPT Hà Nội đã xây dựng được một hệ thống công nghệ thông tin đồ sộ, đáp ứng được đầy đủ các yêu cầu và nhiệm vụ điều hành sản xuất kinh doanh và quản lý trong giai đoạn mới.



Hình 3.1. Các hệ thống thông tin của VNPT Hà Nội

Trong tổng thể các hệ thống thông tin thống nhất đó, hệ thống Phát triển thuê bao (PTTB) nổi lên như một hạt nhân, lưu giữ đầy đủ các thông tin cần thiết và đầy đủ nhất cho bản thân hệ thống Phát triển thuê bao và cho các hệ thống khác hoạt động.

Hệ thống PTTB sở hữu một kho dữ liệu thông tin khổng lồ về khách hàng, thuê bao, khuyến mại.... .

Hiện nay, ngoài VNPT Hà Nội, hệ thống Phát triển thuê bao hiện được Tập đoàn VNPT lựa chọn để triển khai cho 26 tỉnh miền Bắc. Chính vì vậy, lượng dữ liệu của hệ thống hiện đang ngày phát triển thêm.

Trước đây, Cơ sở dữ liệu (CSDL) của hệ thống PTTB được sử dụng phục vụ các tác nghiệp hàng ngày, các báo cáo, thống kê. Những năm gần đây, CSDL của hệ

thống PTTB đã được sử dụng để đáp ứng một phần cho công tác phân tích thông tin.

Trong giai đoạn cạnh tranh khốc liệt nhằm giành thị phần cung cấp dịch vụ viễn thông của các đơn vị như VNPT, Viettel, FPT... thì việc phân tích thông tin để đưa ra các dự đoán về tình hình phát triển thuê bao, dự đoán khả năng thuê bao chấm dứt hợp đồng với các đơn vị cung cấp, phân loại khách hàng để phục vụ cho công tác chăm sóc khách hàng đang rất được quan tâm.

Nghiên cứu lý thuyết khai phá dữ liệu, áp dụng khai phá dữ liệu trên cơ sở dữ liệu khách hàng và thuê bao của hệ thống PTTB VNPT Hà Nội với mong muốn bước đầu tìm hiểu những kết quả khai phá thú vị từ kho thông tin khách hàng và thuê bao VNPT Hà Nội. Những kết quả khai phá dữ liệu trong phạm vi luận văn có thể là bước đầu cho dự án Xây dựng hệ thống phân tích thông tin hỗ trợ các công tác quản lý, phát triển và chăm sóc khác hàng của VNPT Hà Nội.

3.1.2. Mô tả bộ dữ liệu thử nghiệm

Bộ dữ liệu thử nghiệm là bộ dữ liệu của hệ thống Phát triển thuê bao. Dưới đây sẽ mô tả những đối tượng và thuộc tính sẽ được sử dụng để áp dụng vào bài toán phân loại với đối tượng là khách hàng và thuê bao.

Dữ liệu của hệ thống Phát triển thuê bao sau khi xử lý tại module Thu thập dữ liệu và tiền xử lý dữ liệu đã thu được tập thuộc tính sẽ có ảnh hưởng trực tiếp tới việc phân loại đối tượng thuê bao như sau:

- ID Thuê bao: Là dữ liệu chính để phân biệt các thuê bao với nhau. Dữ liệu dạng số và không trùng lặp.
- Số AD: Là mã thuê bao của từng dịch vụ. Ví dụ: Megavnn là số AD (010322918AD), Điện thoại cố định là số điện thoại, MyTV là Account MyTV (HNITV0001527)…
- Loại thiết bị: Loại thiết bị thuê bao sử dụng .Ví dụ : ADSL, GPON, MANE
- Tốc độ: Tốc độ thuê bao : Ví dụ; 12mbps
- Phương thức tính cước: Trọn gói hoặc lưu lượng

- Nhánh: Nhánh của thuê bao(Nhánh chính, nhánh phụ).
- BTS : Thông tin BTS của thuê bao.
- Loại cổng: Loại cổng của thuê bao. Ví dụ: ADSL, SHDSL, STM4, STM16, 34M(16 E1), Cáp quang, Cáp đồng, E1.
- Loại dịch vụ : Loại dịch vụ của thuê bao. Ví dụ: FiberVNN VTN, FiberVNN có xác thực, FiberVNN, Internet trực tiếp với VDC…
- SWITCH : Thông tin Switch của thuê bao.
- Mạng: Thông tin mạng của thuê bao. Ví dụ: VTN1 - Ngan hang Ngoai thuong Viet Nam - Chi Nhanh Thang Long(VFI0000002),Trường đại học nông nghiệp Hà Nội (VFI0000004),VNPT Hà Nội (VFI0000001).
- CIR: Tốc độ cam kết tối thiểu của thuê bao.
- PIR: Tốc độ tối đa của thuê bao.
- Node: Tên node kết nối của thuê bao.
- Kênh: Số lượng kênh thuê của thuê bao.
- Thuê IP: Có thuê IP hay không.
- Số lượng IP thuê: Số lượng IP thuê bao đăng kí thuê.
- Account FTTH: Account của dịch vụ Internet trên đường cáp quang (Với trường hợp có xác thực với hệ thống Visa)
- Loại hình thuê bao: Loại hình của thuê bao. Ví dụ: MyTV đi kèm MegaVNN, Mega đi kèm Điện thoại cố định…
- Đối tương: Đối tượng thuê bao. Ví dụ: Học sinh, sinh viên, Cán bộ y tế, Cán bộ giáo dục…
- Phường xã: Mã phường xã của thuê bao.
- Quận huyện: Mã quận huyện của thuê bao.
- Dịch vụ Viễn thông: Mã dịch vụ của thuê bao. Ví dụ: Megavnn, FTTH, TSL…
- Số lần báo hỏng: Số lần báo hỏng dịch vụ.
- Số lượng tạm dừng yêu cầu: Số lần tạm dừng do yêu cầu của khách hàng

- Số lượng tạm dừng do nợ cước: Số lần tạm dừng do khách hàng bị nợ cước thuê bao.
- Số lượng khôi phục yêu cầu: Số lần khách hàng yêu cầu khôi phục lại dịch vụ đã tạm dừng trước đó.
- Số lượng khôi phục nợ cước: Số lần khôi phục lại dịch vụ đã tạm dừng trước đó do khách hàng đã thanh toán tiền cước thuê bao còn nợ.
- Số lượng thuê bao chung dây: Số lượng các thuê bao trên cùng đường dây.
- Trạng thái thuê bao chung dây: Trạng thái của các thuê bao khác trên cùng đường dây.
- Thời gian sống : Thời gian sống của thuê bao, được tính bằng đơn vị ngày.
- Trạng thái: Trạng thái của thuê bao. Ví dụ : Tháo hủy(TH), Sử dụng (SD).

Dữ liệu tổng hợp liên quan đến danh bạ thuê bao và danh bạ khách hàng đã được gán nhãn theo trạng thái thuê bao và sẽ được sử dụng để xây dựng mô hình phân loại (huấn luyện, kiểm thử).

Để dữ liệu đảm bảo là mới nhất , đáp ứng cho hệ thống thực hiện huấn luyện và kiểm thử thì thay vì sử dụng bảng dữ liệu danh bạ thuê bao ban đầu , luận văn đề xuất việc xây dựng các Job thực hiện tự động tổng hợp và lấy các thông tin từ các bảng có liên quan đến thuê bao như: danh bạ thuê bao, hợp đồng tạm dừng/ khôi phục, bảng dữ liệu yêu cầu sửa chữa từ hệ thống điều hành sửa chữa,… Các Job tổng hợp dữ liệu có thể chạy tự động và có thể điều chỉnh được thời gian lấy dữ liệu theo yêu cầu.

Dữ liệu được sử dụng trong bước áp dụng mô hình là dữ liệu danh sách thuê bao hiện đang tạm dừng và được chuẩn hóa như dạng dữ liệu được sử dụng trong quá trình học, điểm khác đó là bảng dữ liệu này sẽ chưa được gán nhãn và sẽ được gán nhãn sau khi áp dụng mô hình.

Dữ liệu sau khi áp dụng mô hình sẽ được lưu vào một bảng gồm có các thông tin như: Thuê bao ID, Trạng thái và tỷ lệ dự đoán đúng (nếu có). Các hệ thống khác sẽ sử dụng dữ liệu này để xây dựng các yêu cầu liên quan hoặc sẽ được tổng hợp lại tại module thu thập dữ liệu để áp dụng cho lần chạy phân loại tiếp theo.

3.2. Cài đặt và thử nghiệm

3.2.1. Yêu cầu về phần cứng thử nghiệm

Để đáp ứng việc thử nghiệm đảm bảo tốc độ phân loại, luận văn đề xuất yêu cầu về phần cứng thử nghiệm tối thiểu như sau:

Hệ điều hành: Microsoft Windows 7

Bộ xử lý: Intel Core i3 4130 @ 3.40GHz.

Bộ nhớ Ram: 3GB

3.2.2. Yêu cầu về công cụ và phần mềm sử dụng

Cơ sở dữ liệu (CSDL) sử dụng trong luân văn lấy từ CSDL của hệ thống Phát triển thuê bao của VNPT Hà Nội, hệ thống này sử dụng CSDL Oracle. Do vậy việc chọn công cụ khai phá dữ liệu của hãng Oracle cũng là một lựa chọn tất yếu.

Khai phá dữ liệu bằng sản phẩm của hãng Oracle, có thể lựa chọn một trong 2 ứng dụng sau:

Darwin: Là một ứng dụng khai phá dữ liệu đặc biệt để xử lý với nhiều gigabytes dữ liệu và cung cấp những câu trả lời cho các bài toán phức tạp như phân lớp dữ liệu, dự đoán và dự báo. Phần mềm Darwin giúp ta chuyển đổi một khối lượng dữ liệu lớn thành những tri thức kinh doanh (tri thức nghiệp vụ - Business intelligence). Darwin giúp tìm ra những mẫu và các liên kết có ý nghĩa trong toàn bộ dữ liệu – Các mẫu cho phép ta hiểu tốt hơn và dự đoán được hành vi của khách hàng.

Oracle Data Mining (ODM) được thiết kế cho người lập trình, những nhà phân tích hệ thống, các quản trị dự án và cho tất cả những ai quan tâm đến việc phát triển các ứng dụng CSDL dùng khai phá dữ liệu để phát hiện ra các mẫu ẩn và dùng tri thức đó để tạo các dự đoán. ODM là công cụ khai phá dữ liệu được nhúng trong CSDL Oracle. Dữ liệu không tách rời CSDL - dữ liệu, và tất cả những hoạt động chuẩn bị dữ liệu, xây dựng mô hình và áp dụng mô hình đều được giữ trong CSDL. Việc này cho phép Oracle xây dựng nền tảng cho những nhà phân tích dữ liệu và những người phát triển ứng dụng có thể tích hợp khai phá dữ liệu một cách liền mạch với các ứng dụng CSDL.

Oracle Data Mining cung cấp các chức năng dự đoán như cho trên Bảng 3.1 và chức năng mô tả trên Bảng 3.2.

Bảng 3.1. Các chức năng dự đoán của ODM

Chức năng	Mô tả	Các thuật toán
Phân lớp Classification	Mô hình phân lớp dùng dữ liệu lịch sử để dự đoán dữ liệu rời rạc hoặc phân loại mới	Naive Bayes, Adaptive Bayes Network, Support Vector Machine, Decision Tree
Phát hiện bất thường Anomaly Detection	Mô hình phát hiện bất thường dự đoán có hay không một điểm dữ liệu là điển hình cho sự phân tán cho trước. PL/SQL và Java APIs hỗ trợ phát hiện bất thường qua chức năng phân lớp	One-Class Support Vector Machine (SVM). PL/SQL và Java APIs hỗ trợ One-Class SVM dùng chức năng khai phá phân lớp và thuật toán SVM không có đích.
Hồi qui Regression	Mô hình Hồi qui dùng dữ liệu lịch sử để dự đoán dữ liệu số, liên tiếp mới	Support Vector Machine
Độ quan trọng của thuộc tính Attribute Importance	Mô hình độ quan trọng của thuộc tính xác định tầm quan trọng liên quan của một thuộc tính trong việc dự đoán một đầu ra cho trước.	Minimal Descriptor Length

Bảng 3.2. Các chức năng mô tả của ODM

Chức năng	Mô tả	Các thuật toán
Phân nhóm Clustering	Mô hình phân nhóm xác định các nhóm tự nhiên trong tập dữ liệu	Enhanced k-means, Orthogonal Clustering (O-Cluster - Thuật toán bản quyền của Oracle)
Các luật kết hợp Association Rules	Mô hình kết hợp xác định các quan hệ và khả năng xuất hiện của chúng trong tập dữ liệu	Apriori
Trích chọn đặc trưng Feature Extraction	Mô hình trích chọn đặc trưng tạo tập dữ liệu tối ưu làm cơ sở cho mô hình trên đó.	Non-Negative Matric Factorization

Một trong những công cụ khá trực quan và dễ thiết lập cho việc khai phá dữ liệu Oracle đó là công cụ Oracle SQL Developer (Phiên bản 4.1.0 được cung cấp miễn phí tại địa chỉ: http://www.oracle.com/technetwork/developer-tools/sql-developer/downloads/index.html) . Phiên bản này thích hợp với Database 11g trở lên.

- **Mô tả phương pháp Naïve Bayes của ODM**

  Huấn luyện

Đầu vào:

Tập dữ liệu với các đối tượng đã được gán nhãn và các thuộc tính của đối tượng. Có 2 phân lớp có thể: c_1 (“TH”) và c_2 (“SD”), trong đó TH là lớp khách hàng đã tháo hủy dịch vụ và SD là lớp khách hàng đang sử dụng dịch vụ.

Đầu ra:

Mô hình Naïve Bayes (Các giá trị xác suất cho mỗi phân lớp $P(c_1)$, $P(c_2)$; Giá trị xác suất xảy ra của giá trị thuộc tính x_j đối với một phân lớp c_i: $P(x_j \mid c_i)$)

Các số liệu về tỷ lệ phân loại của các nhãn "SD" và "TH"

Quá trình học Bayes đơn giản là quá trình tính các xác suất $P(c_j)$ và các xác suất điều kiện $P(x_i|c_j)$ bằng cách đếm trên tập dữ liệu.

Phân loại

Đầu vào:

Tập dữ liệu chưa được phân lớp

Mô hình NB

Đầu ra:

Nhãn/lớp của đối tượng thuê bao cần phân loại .

Giai đoạn phân loại mô hình NB xử lý các công việc sau:

Tính toán xác suất có thể xảy ra (likelihood) của ví dụ z đối với mỗi phân lớp

- Đối với phân lớp c_1 ("TH") : $P(z|c_1)$
- Đối với phân lớp c_2 ("SD"): $P(z|\ c_2)$

Xác định phân lớp có thể nhất (the most probable class)

- Đối với phân lớp c_1: $P(c_1).P(z|\ c_1)$
- Đối với phân lớp c_2: $P(c_2).P(z|\ c_2)$

So sánh $P(c_1).P(z|\ c_1)$ với $P(c_2).P(z|\ c_2)$ để gán nhãn cho tập z

- **Mô tả phương pháp Support Vector Machine của ODM**

- Tiền xử lý dữ liệu

Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector số thực, ta cần phải tìm cách chuyển về dạng số của SVM.

Thực hiện chuyển đổi dữ liệu sao phù hợp cho quá trình tính toán, thường co giãn dữ liệu để chuyển về đoạn [-1,1] hoặc [0,1].

Trong tập dữ liệu cho trước, giả sử mỗi đối tượng được biểu diễn bởi bộ gồm n thuộc tính đơn (a_1, a_2,….,a_n), mỗi thuộc tính a_k đó có tương ứng m_k trạng thái. Khi đó, để sử dụng chúng ta sẽ chuyển đổi dữ liệu sang dạng của SVM. Tổng tất cả các thuộc tính sẽ là : $M = m_1 + m_2 + \ldots + m_n$, được xếp lần lượt từ 1 đến m. Vì vậy ta sẽ dùng vectơ nhị phân biểu diễn gồm m ký tự 0/1, trong đó: 1 tương ứng với thuộc

tính được chọn cho đối tượng đó và 0 với trường hợp không có thuộc tính này. Ta có thể rút gọn cách biểu diễn này nhờ vào cách biểu diễn [vị trí:giá trị].

Ví dụ: Có m vị trí thuộc tính đã xác định trước sẽ được biểu diễn rút gọn như sau: -1 2:1 5:1 7:1 15:1 22:1 35:1 60:1 102:1 112:1

Trong đó -1 biểu thị tên lớp, có các thuộc tính tương ứng ở vị trí 2, 5, 7, 15, 22, 35, 60, 102, 112.

- ***Huấn luyện và kiểm thử tập dữ liệu kiểm tra***

Đầu vào:

Các vector đặc trưng biểu diễn tập dữ liệu các đối tượng thuê bao trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector) và tập kiểm thử.

Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.

Các tham số cho mô hình SVM (nếu có): C, γ (tham số của hàm kernel, thường dùng hàm Gauss hoặc Linear)

Đầu ra:

Mô hình SVM (Các Support Vector, w, tham số b). Các số liệu về tỷ lệ phân loại của các nhãn "SD" và "TH"

- ***Phân loại***

Đầu vào:

Vector đặc trưng của đối tượng thuê bao mới cần phân lớp.

Mô hình SVM

Đầu ra:

Nhãn/lớp của đối tượng thuê bao cần phân loại.

- **Thiết lập và chạy thử nghiệm công cụ Oracle Data Miner**

Sau đây xin giới thiệu các bước thiết lập Oracle Data Miner bằng công cụ Oracle SQL Developer 4.1.0.

Bước 1: Tạo một Account Data Miner.

- Tạo mới một account để khai phá dữ liệu .
- Tạo kết nối cho User Data Miner.

Bước 2: Cài đặt Data Miner.

- Tạo một Project Data Miner.
- Tạo một Data Miner Workflow.

Bước 3 : Tạo phân lớp Classification

Vì hệ thống muốn dự đoán các khách hàng hoặc thuê bao có nhiều khả năng tháo hủy dịch vụ. Do đó, luận văn sẽ chỉ định mô hình phân loại. Theo mặc định, Oracle Data Miner chọn tất cả các thuật toán hỗ trợ cho một mô hình phân loại. Ở đây luận văn chỉ lựa chọn 02 thuật toán là Naïve Bayes và SVM.

Bước 4: Áp dụng mô hình

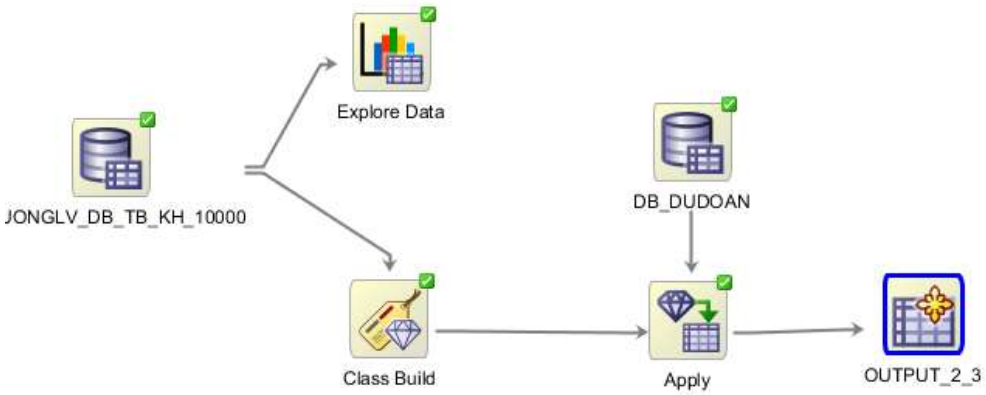
Phần này sẽ áp dụng mô hình Naïve Bayes (NB) hoặc SVM đã được đánh giá độ chính xác và tạo ra một bảng để hiển thị kết quả , dùng để dự đoán trường hợp khách hàng hoặc thuê bao có thể thực hiện yêu cầu tháo hủy dịch vụ.

Trước khi ta thực hiện cần xem xét các kết quả đầu ra. Theo mặc định, chương trình sẽ tạo ra hai cột thông tin cho mỗi thuê bao của khách hàng:

- Dự đoán (“TH” hoặc “SD”)
- Xác suất dự đoán

Tuy nhiên, ta thực sự muốn biết thông tin này cho mỗi thuê bao, để dễ dàng kết hợp các thông tin với một thuê bao nhất định. Để có được thông tin này cần phải thêm một cột thứ ba của dữ liệu đầu ra là Thuebao_ID.

Kết thúc tất cả các bước sẽ có mô hình được biểu diễn như hình 3.2.



Hình 3.2 Mô hình của hệ thống ODM đã thiết lập trên công cụ SQL Developer

3.3. Kết quả và đánh giá

Để đánh giá tính đúng (tương đối) của một mô hình, cần sử dụng các tập mẫu dữ liệu Ω mà mỗi mẫu được biết trước thuộc về một trong K lớp: c_1, c_2,…,c_k.

Mặt khác tập Ω được phân hoạch thành hai tập con $\Omega = \Omega_{Tr} \cup \Omega_{Ts}$,

$\Omega_{Tr} \cap \Omega_{Ts} = \varnothing$ trong đó :

Ω_{Tr}: Tập mẫu dữ liệu dùng để huấn luyện trong quá trình học.

Ω_{Ts}: Tập mẫu dữ liệu dùng để kiểm tra (Test), đánh giá tính đúng của mô hình.

Trong khuôn khổ luận văn này, dữ liệu sử dụng đã được xác định. Từ tập dữ liệu về thuê bao và khách hàng này, ta sẽ thực hiện phân tách thành 2 tập dữ liệu con. Để thuận tiện cho quá trình đánh giá giải thuật phân lớp, ta sẽ tách theo các trường hợp với tỷ lệ tương ứng như sau:

- Ω_{Tr} = 90% Ω, Ω_{Ts}= 10% Ω
- Ω_{Tr} = 80% Ω, Ω_{Ts}= 20% Ω
- Ω_{Tr} = 70% Ω, Ω_{Ts}= 30% Ω
- Ω_{Tr} = 60% Ω, Ω_{Ts}= 40% Ω
- Ω_{Tr} = 50% Ω, Ω_{Ts}= 50% Ω

Một số chỉ số thông dụng được dùng để đánh giá giải thuật SVM và Naïve Bayes, hay cụ thể là để đánh giá một bộ phân loại hai lớp “TH” và “SD” là:

- Số đúng TH(**TP**-True positive): Số phần tử nhãn TH được phân loại TH.
- Số sai TH (**FP** – False positive): Số phần tử có nhãn TH được phân loại SD
- Số đúng SD(**TN** –True negative) : Số phần tử có nhãn SD được phân loại SD.
- Số sai SD(**FN** – False negative): Số phần tử có nhãn SD được phân loại TH.
- Tỷ lệ phân loại đúng: Phần trăm thuê bao được phân loại đúng nhãn không phân biệt đó là nhãn “TH” hay “SD” .

Tỷ lệ phân loại đúng = (Tỷ lệ phân loại đúng ”TH” + Tỷ lệ phân loại đúng “SD”) / 2

- Tỷ lệ phân loại sai: 100% - Tỷ lệ phân loại đúng

Bảng 3.3 thống kê kết quả khi chạy thử nghiệm mô hình phân loại với tập dữ liệu thuê bao và khách hàng của VNPT Hà Nội. Tổng số bản ghi tương ứng với số thuê bao là : 2,674,663 bản ghi.

Bảng 3.3. Thống kê kết quả thử nghiệm với thuật toán Naïve Bayes và SVM

Lần huấn luyện	Tỷ lệ tập huấn luyện (%)	Tỷ lệ tập kiểm thử (%)	Thuật toán Naïve Bayes		Thuật toán SVM(Hàm nhân Linear)	
			Tỷ lệ phân loại đúng (%)	Tỷ lệ phân loại sai (%)	Tỷ lệ phân loại đúng (%)	Tỷ lệ phân loại sai (%)
Lần 1	90	10	98,167	1,833	98,491	1,509
Lần 2	80	20	98,169	1,831	98,455	1,545
Lần 3	70	30	98,154	1,846	98,422	1,578
Lần 4	60	40	98,164	1,836	98,424	1,576
Lần 5	50	50	98,167	1,833	98,359	1,641

Theo bảng 3.3 , qua 5 lần thử nghiệm đối với mỗi thuật toán ta có các kết luận sau:

- Thuật toán SVM với hàm nhân Linear cho kết quả phân loại với tỉ lệ phân loại đúng tốt nhất là 98.491% với trường hợp tỷ lệ tập huấn luyện và tập kiểm thử là 90% và 10%.
- Thuật toán Support Vector Machine (SVM) với hàm nhân Linear có độ chính xác trong phân lớp cao hơn Thuật toán Naïve Bayes (NV).

Như vậy theo kết quả thực nghiệm cho thấy với dữ liệu thử nghiệm như trên phương pháp SVM cho kết quả tốt hơn so với phương pháp Naïve Bayes . Tuy nhiên , phương pháp Bayes có ưu thế về tốc độ phân loại do có độ phức tạp tính toán thấp hơn trong khi SVM đòi hỏi khối lượng và thời gian tính toán lớn hơn. Theo thống kê , với cùng tỷ lệ tập huấn luyện và kiểm thử là 90% và 10%, thời gian

để thuật toán Naïve Bayes thực hiện cả 2 bước huấn luyện và kiểm thừ là 17 phút còn thuật toán SVM là 20 phút.

Luận văn đã xác định được mô hình phân loại tốt nhất, đáp ứng được yêu cầu mà bài toán phân loại đã đặt ra. Mục tiêu thử nghiệm tiếp theo sẽ đưa ra đánh giá về ảnh hưởng của từng thuộc tính tới quyết định và hành vi của khách hàng. Các đánh giá này sẽ giúp ích rất nhiều cho phương hướng kinh doanh của VNPT Hà Nội, và giúp cho việc hoạch định các chính sách với từng nhóm thuê bao có khả năng tháo hủy cao.

Luận văn sẽ lấy trường hợp tập kiểm thử chiếm 10% để thử nghiệm vì trường hợp này cho kết quả phân loại đúng cao nhất và luận văn cũng chỉ thử nghiệm bằng thuật toán Naïve Bayes. Kết quả thực nghiệm với 7 trường hợp loại bỏ thuộc tính được biểu diễn tại bảng 3.4 .

Bảng 3.4. Kết quả thử nghiệm xác định ảnh hưởng thuộc tính

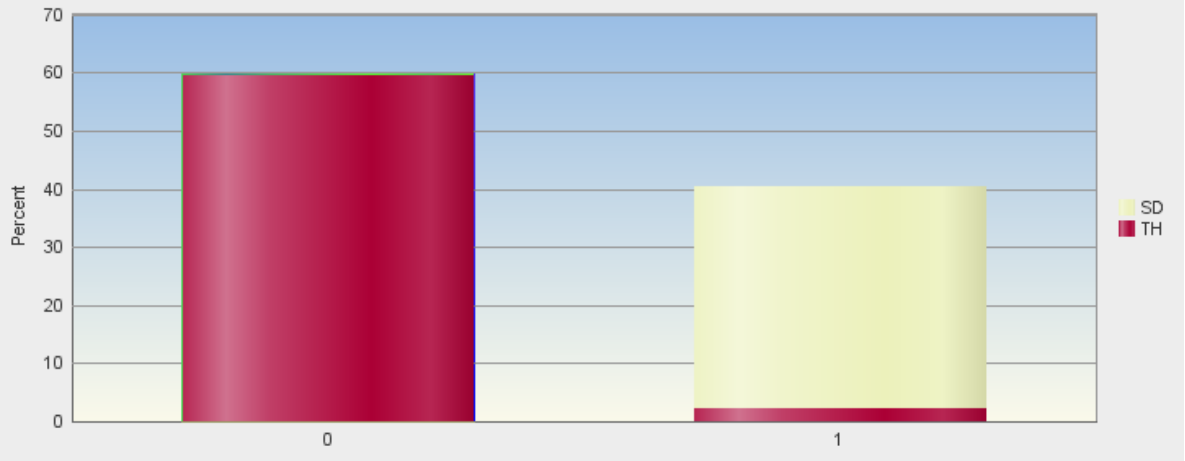
Thuộc tính được loại bỏ	**Tỷ lệ phân loại đúng với dữ liệu đã loại bỏ thuộc tính (%)**
Thời gian sống của thuê bao	98.0394
Trạng thái thuê bao cùng dây	81.6493
Số lần báo hỏng	98.1444
Dịch vụ viễn thông của thuê bao	98.1039
Phương thức tính cước	98.1982
Đối tượng thuê bao	97.9218
Địa chỉ thuê bao(phường xã)	98.2167

Với số liệu thử nghiệm có được tại bảng 3.4, thuộc tính Trạng thái của thuê bao cùng dây có ảnh hưởng nhiều nhất tới độ chính xác của mô hình phân loại, thuộc tính về địa chỉ thuê bao (phường xã) có ảnh hưởng ít nhất. Để xác định cụ thể

ảnh hưởng chi tiết của thuộc tính này, luận văn sử dụng chức năng Explore Data của công cụ thử nghiệm.

Data Miner tính toán một loạt các thông tin về từng thuộc tính trong dữ liệu, vì nó liên quan đến thuộc tính đã xác định, bao gồm các giá trị biểu đồ ,trung bình, Min và Max giá trị, sai lệch chuẩn, phương sai,…

Màn hình hiển thị cho phép hình dung và xác nhận các dữ liệu, và cũng có thể tự kiểm tra các dữ liệu. Hình 3.4 mô tả chi tiết ảnh hưởng của thuộc tính tới phân loại.



Hình 3.3. Mô tả thuộc tính của Data Source trong Data Miner

Theo như hình 3.3:

- 59.71% Thuê bao có thuê bao cùng dây ở trạng thái tháo hủy sẽ tháo hủy theo.

- 2.27% Thuê bao có thuê bao cùng dây ở trạng thái sử dụng sẽ tháo hủy.

- 38.02% Thuê bao có thuê bao cùng dây ở trạng thái vẫn sử dụng sẽ vẫn duy trì sử dụng dịch vụ

- Tại module Áp dụng mô hình:

Kết thúc việc thực hiện quá trình Áp dụng mô hình , dữ liệu thông tin dự đoán các thuê bao có thể thực hiện tháo hủy được lưu vào một bảng dữ liệu với khuôn dạng như bảng 3.5 .

Bảng 3.5. Dữ liệu thông tin dự đoán khách hàng tháo hủy dịch vụ

THUEBAO ID	TRẠNG THÁI	XÁC SUẤT DỰ ĐOÁN
15729	TH	92.46
14126	TH	91.13
36110	TH	91.19
1383	TH	91.18
6620	TH	92.23
3591	TH	91.16
19842	TH	91.18
25905	TH	96.09
8700	TH	91.40
32130	TH	96.19

Thông tin trạng thái và xác suất dự đoán sẽ được ODM tự động tính toán dựa trên dữ liệu dự đoán và mô hình dự đoán đã chọn. Hàm tính xác suất dự đoán : PREDICTION_PROBABILITY([schema.]model [,class] mining_attribute_clause)

Dữ liệu dự đoán sau bước áp dụng mô hình sẽ được sử dụng làm dữ liệu đầu vào cho các đơn vị lấy số liệu thống kê hoặc các hệ thống khác khi cần. Dữ liệu này sẽ rất cần thiết cho phòng kế hoạch kinh đoanh dưa ra chỉ tiêu phát triển thuê bao cho trung tâm kinh doanh, phòng mạng và dịch vụ đưa ra chỉ tiêu thiết lập dịch vụ cho các trung tâm viễn thông…

3.4. Kết chương

Chương III giới thiệu về bộ dữ liệu thử nghiệm, xây dựng mô hình, cài đặt thực nghiệm và đánh giá mô hình phân loại khách hàng trên bộ dữ liệu thử nghiệm của hệ thống PTTB đang triển khai tại VNPT Hà Nội dựa trên các thuật toán học máy có giám sát. Kết quả thực nghiệm khẳng định thuật toán Support Vector cho kết quả phân loại tốt hơn thuật toán Machine Naïve Bayes. Mô hình phân loại và dự đoán hành vi của khách hàng và thuê bao của VNPT Hà Nội dựa trên bộ công cụ Oracle Data Miner và công cụ SQL Developer đơn giản, dễ cài đặt và cho kết quả rất thống kê và so sánh rất chi tiết.

KẾT LUẬN

1. Những đóng góp của luận văn

Với mục tiêu nghiên cứu và xây dựng mô hình dự đoán hành vi của khách hàng và thuê bao tại VNPT Hà Nội,luận văn đã đi sâu nghiên cứu các phương pháp học máy có giám sát và ứng dụng, thực hiện xây dựng mô hình và thử nghiệm các giải thuật học máy Naïve Bayes và Support Vector Machine (SVM) để phân loại khách hàng.

Những kết quả chính đã đạt được trong luận văn:

- Nghiên cứu tổng quan một số vấn đề về học máy, học máy có giám sát bao gồm ứng dụng và một số thuật toán học máy áp dụng vào bài toán phân loại, trong đó chú trọng các phương pháp học máy có giám sát. Ngoài ra, luận văn cũng giới thiệu được tổng quan về bài toán phân loại và dự đoán dựa trên cơ sở dữ liệu khách hàng và thuê bao của VNPT Hà Nội.
- Nghiên cứu sâu hai thuật toán phân loại học máy có giám sát là Naïve Bayes và SVM; từ đó đưa ra bài toán áp dụng vào phân loại khách hàng và dự đoán hành vi khách hàng.
- Xây dựng mô hình, thực nghiệm và đánh giá kết quả phân loại dựa trên các thuật toán học máy có giám sát. Kết quả thực nghiệm khẳng định thuật toán Naïve Bayes cho kết quả phân loại tương đối tốt, thời gian thực hiện nhanh.

2. Hướng phát triển của luận văn

Kết quả nghiên cứu đã đạt được của luận văn có thể làm cơ sở vững chắc cho việc tích hợp tính năng phân loại và dự đoán hành vi khách hàng vào các hệ thống dịch vụ của VNPT Hà Nội như Phát triển thuê bao, Chăm sóc khách hàng, Tính cước...

Hiện nay, hệ thống Phát triển thuê bao của VNPT Hà Nội đã và đang triển khai cho VNPT các tỉnh miền Bắc, vì vậy cơ sở dữ liệu khách hàng và thuê bao sẽ rất lớn. Đó cũng là tiền đề để ta có thể áp dụng mô hình phân loại và dự đoán đã thử nghiệm trên bộ cơ sở dữ liệu lớn hơn.

DANH MỤC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Trần Cao Đệ, Phạm Nguyên Khang (2012), *Phân loại văn bản với máy học vector hỗ trợ và cây quyết định*, 52-63.

[2] Từ Minh Phương (2010), *Nhập môn trí tuệ nhân tạo*, Học Viện Công Nghệ Bưu Chính Viễn Thông, Hà Nội.

[3] Võ Văn Tài (2012), *Phân loại bằng phương pháp Bayes từ số liệu rời rạc*, 69-78.

[4] Phùng thị Anh (2014) , *Một số phương pháp phân lớp dữ liệu và ứng dụng phận lớp dịch vụ Web*.

Tiếng Anh

[5] Xindong Wu , Vipin Kumar (2009) , *The Top Ten Algorithms in Data Mining*.

[6] TAlberto Tellaechea, Xavier P. Burgos-Artizzub, Gonzalo Pajaresa, Angela Ribeirob (2008), *Avision-basedmethod forweeds identification through the Bayesian decision theory*, 521-530.

[7] Burges C (1998), *A tutorial on Support Vector Machines for pattern recognition, Proceedings of Int Conference on Data Mining and Knowledge Discovery*, 121-167.

[8] Chang, C.C., Lin, C.J (2011), *LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology* .

[9] Robert C. P (2001), *The Bayesian Choice*, Springer-Verlag.

[10] ZhaoHui Tang, Jamie MacLennan (2005), *Data Mining with SQL Server 2005*.

[11] Oracle (2008), *Data Mining Application Developer's Guide*, 100 trang.

Các trang Web tham khảo

[12] Oracle, *Text Mining Using Oracle Data Mining,* http://docs.oracle.com/cd/B14117_01/datamine.101/b10698/8text.htm, truy cập 11/2014

[13] Nguyễn Văn Chức, *Giới thiệu công cụ xây dựng mô hình khai phá dữ liệu BIDS của Microsoft,* http://bis.net.vn/forums/t/458.aspx, truy cập 11/2014