Trường Đại Học Bách Khoa Hà Nội
Viện Toán ứng dụng và Tin học

BÁO CÁO MÔN HỌC MÔ HÌNH NGẪU NHIÊN VÀ ỨNG DỤNG

Đề tài: PHÁT HIỆN GIAN LẬN THẺ TÍN DỤNG BẰNG MÔ HÌNH MARKOV ẨN

Giảng viên HD: TS. NGUYỄN THỊ NGỌC ANH

Sinh viên: NGUYỄN THANH HẢI - 20170734
NGUYỄN TIẾN THÀNH - 20173374
PHẠM TIẾN DUẬT - 20173497

Lớp: CTTN-TOÁN TIN K62

Hà Nội, tháng 7 năm 2020

Mục lục

Lời nói đầu **2**

1 Giới thiệu **4**

2 Một số định nghĩa mở đầu và Mô hình Markov **6**
2.1 Một số định nghĩa 6
2.2 Mô hình Markov 7

3 Mô hình Markov ẩn **9**
3.1 Mô hình Markov ẩn 9
3.1.1 Khái niệm 9
3.1.2 Ví dụ 10
3.2 Các thành phần và hoạt động của mô hình Markov ẩn 13
3.2.1 Các thành phần của mô hình Markov ẩn 13
3.2.2 Hoạt động của mô hình Markov ẩn 14
3.2.3 Các công thức được sử dụng trong mô hình Markov ẩn 14
3.3 Ba bài toán cơ bản của mô hình Markov ẩn 15
3.3.1 Bài toán 1: Bài toán ước lượng (Likelihood Computation) 15
3.3.2 Bài toán 2: Bài toán giải mã (Decoding Computation) 16
3.3.3 Bài toán 3: Bài toán huấn luyện (Training HMM) 19
3.4 Một số ứng dụng của mô hình Markov ẩn 20

4 Phát hiện gian lận sử dụng mô hình Markov ẩn **22**
4.1 Phát hiện gian lận sử dụng mô hình Markov ẩn 22
4.2 Áp dụng phương pháp 24
4.2.1 Thuật toán phân cụm 24
4.2.2 Thuật toán đào tạo 24
4.2.3 Thuật toán phát hiện 26
4.3 Nhận xét 26
4.3.1 Hiệu quả của phương pháp 26
4.3.2 Cách khắc phục và phương hướng phát triển 27

Kết luận **29**

Tài liệu tham khảo **30**

Lời nói đầu

Hiện nay, xã hội đang phát triển không ngừng đặc biệt là các thành tựu về khoa học - kĩ thuật. Trong thời đại công nghiệp 4.0, các nền tảng công nghệ số đang phát triển rất mạnh mẽ và góp phần không nhỏ trong đó là các kết quả ứng dụng nghiên cứu của bộ môn Toán – Tin ứng dụng, và đặc biệt đó là các kết quả của lĩnh vực xác suất – thống kê. Lĩnh vực này đã có rất nhiều thành tựu trong nghiên cứu cũng như thực tiễn, trong đó có những thành tựu đến từ các mô hình thống kê.

Mô hình Markov ẩn (Hidden Markov Model - HMM) là mô hình thống kê trong đó hệ thống được mô hình hóa được cho là một quá trình Markov với các tham số không biết trước. Nhiệm vụ là xác định các tham số ẩn từ các tham số quan sát được, dựa trên sự thừa nhận này. các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp. Mô hình HMM đã có rất nhiều những ứng dụng trong thực tế như dự báo thời tiết, nhận dạng tiếng nói, xử lí tín hiệu... Và gần đây, thực tế đã có những bước đi đầu tiên trong việc nghiên cứu ứng dụng của mô hình Markov ẩn vào phát hiện gian lận thẻ tín dụng.

Trên cơ sở gợi ý, hướng dẫn của T.S Nguyễn Thị Ngọc Anh, chúng em đã thực hiện bài báo cáo với nội dung tập trung nghiên cứu về lý thuyết của mô hình Markov ẩn và ứng dụng của mô hình Makov ẩn vào bài toán phát hiện gian lận thẻ tín dụng. Báo cáo ngoài phần mục lục, lời nói đầu, kết luận và tài liệu tham khảo thì bao gồm 4 chương:

Chương 1: Giới thiệu: trình bày những vấn đề trong thực tế dẫn tới bài toán "phát hiện gian lận thẻ tín dụng" và giới thiệu một số phương pháp.

Chương 2: Một số định nghĩa khởi đầu và mô hình markov. Trình bày cơ sở lý thuyết chuẩn bị cho mô hình Markov ẩn.

Chương 3: Mô hình Markov ẩn. Chương này trình bày nội dung lý thuyết của mô hình Markov ẩn, những bài toán cơ bản và một số ứng dụng trong thực tế.

Chương 4: Phát hiện gian lận sử dụng mô hình Markov ẩn. Chương này trình bày cơ sở lý thuyết của phương pháp, trình bày các thuật toán sử dụng để giải quyết bài toán "phát hiện gian lận thẻ tín dụng sử dụng mô hình Markov ẩn". Sau đó là phần nhận xét của phương pháp cũng như những hạn chế và hướng phát triển của phương pháp.

Chúng em xin gửi lời cảm ơn sâu sắc đến cô – TS. Nguyễn Thị Ngọc Anh cùng các bạn trong lớp CTTN Toán Tin K62 đã có những góp ý quý báu giúp bài báo cáo được hoàn thành.

Mặc dù đã có nhiều cố gắng trong quá trình thực hiện nhưng báo cáo chắc chắn không tránh khỏi thiếu sót. Vì vậy, chúng em rất mong nhận được những góp ý của thầy cô cùng các bạn để bài báo cáo được hoàn thiện hơn.

Hà Nội, tháng 07/2020

Nhóm sinh viên thực hiện

Chương 1

Giới thiệu

Ngày nay, các phương thức thanh toán trực tuyến xuất hiện ngày càng nhiều và dần thay thế tiền mặt. Các ngân hàng cung cấp các hình thức thanh toán khác nhau như tiền ảo, thẻ tín dụng, ngân hàng trực tuyến, ví điện tử... để phục vụ nhu cầu của thị trường. Trong đó, thẻ tín dụng là một trong những hình thức giao dịch phổ biến nhất. Thẻ tín dụng được phân thành hai loại:

- Thẻ vậy lý. Thẻ vật lý dùng để thanh toán tại các cửa hàng (quẹt thẻ) hoặc thanh toán trực tuyến.
- Thẻ điện tử. Thẻ điện tử không có ở dạng vật lý, nó là một công cụ thanh toán kỹ thuật số.

Thẻ tín dụng (Credit Card) là một loại thẻ ngân hàng mà người sở hữu có thể dùng để thanh toán mà không cần tiền có sẵn trong thẻ. Điều này có nghĩa là bạn "mượn" một số tiền của ngân hàng để mua sắm, chi tiêu và cuối kỳ sẽ phải trả lại đầy đủ cho ngân hàng. Thẻ tín dụng được làm bằng chất liệu nhựa polyme với hình dạng và kích thước theo tiêu chuẩn ISO 7810. Tùy vào ngân hàng phát hành mà thẻ sẽ có màu sắc cùng thiết kế riêng biệt.

Gian lận thẻ tín dụng là hình thức gian lận sử dụng công nghệ cao để đánh cắp thông tin thẻ tín dụng (Visa, MasterCad, ATM...) của người sử dụng thuộc về lĩnh vực tài chính, ngân hàng. Các hình thức gian lận:

- Bị thanh toán hay quẹt thẻ tại một cửa hàng nào đó khi mua hàng trong trường hợp bạn bị trộm thẻ.
- Sử dụng công nghệ cao qua mạng Internet đánh cắp thông tin thẻ tín dụng của người dùng.
- Bị rút trộm tiền mặt qua máy ATM.
- Làm giả thẻ Visa, MasterCad.

Thiệt hại từ các vụ gian lận thẻ tín dụng là rất lớn. Theo Nilson Report, số lượng giao dịch bằng thẻ tín dụng trên toàn thế giới ngày càng tăng, đến năm 2016 có đến 257,17 tỷ giao dịch liên quan đến thẻ tín dụng [1]. Chính vì khách hàng ngày càng ưa chuộng hình thức thanh toán qua thẻ tín dụng nên một số vụ lừa đảo, gian lận liên quan đến loại hình thanh toán này có xu hướng tăng. Theo số liệu của Nilson, năm 2015 cả thế giới bị mất 28,8 tỷ đô là do các vụ lừa

đảo hoặc gian lận về thẻ tín dụng. Trong đó số vụ gian lận ở Châu Á Thái Bình dương là lớn nhất [2]. Vì vậy, các ngân hàng đã phải xây dựng các hệ thống ngăn ngừa và phát hiện gian lận trong các giao dịch, trong đó có giao dịch thẻ tín dụng. Hệ thống này hoạt động theo quy tắc chung:

- Lưu trữ thông tin giao dịch của khách hàng.
- Phân tích dữ liệu giao dịch và khoanh vùng phạm vi sử dụng thẻ.
- Khi một giao dịch được tiến hành, hệ thống dựa vào dữ liệu trong quá khứ để phỏng đoán xem giao dịch có gì bất thường không, từ đó đưa ra các cảnh báo cần thiết.

Các nhà nghiên cứ đã đưa ra nhiều kỹ thuật để phát hiện gian lân thẻ tín dụng như: phát hiện gian lận dựa vào cây quyết định [3]; sử dụng mạng Nơ ron để phán đoán các giao dịch bất thường [4]; kỹ thuật sử dụng cây quyết đinh cite5... Trong báo cáo này, chúng em giới thiệu và tập trung phân tích kỹ thuật sử dụng mô hình Markov ẩn (Hidden Markov Model – HMM) để phát hiện gian lận thẻ tín dụng.

Chương 2

Một số định nghĩa mở đầu và Mô hình Markov

2.1 Một số định nghĩa

Định nghĩa 2.1: *(Tính Markov). Một dãy trạng thái ngẫu nhiên được gọi là có tính Markov nếu xác suất chuyển trạng thái tiếp theo chỉ phụ thuộc vào trạng thái hiện tại và quá khứ.*

Định nghĩa 2.2: *(Tính thuần nhất). Một trạng thái ngẫu nhiên được gọi là có tính thuần nhất nếu xác suất chuyển trạng thái chỉ phụ thuôc vào độ lớn khoảng thời gian mà không phụ thuộc vào thời điểm.*

Định nghĩa 2.3: *(Xích Markov). Dãy biến ngẫu nhiên $(X_n)_{n\geq 0}$ được gọi là một xích Markov với phân phối ban đầu λ và ma trận chuyển P nếu:*

i, *X_0 có phân phối λ tức là:*

$$P(X_0 = i) = \lambda_i \qquad \forall i \in I$$

ii, *Với mọi $n \geq 0$ phân phối của X_{n+1} với điều kiện $X_n = i_n$ là $(P_{i_n j})_{j\in I}$ và độc lập với $X_0, \ldots, X_{n-1}$. Tức là:*

$$\begin{aligned} P(X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_0 = i_0) &= P(X_{n+1} = i_{n+1} \mid X_n = i_n) \\ &= P_{i_n i_{n+1}} \qquad \forall n \geq 0, i_0, \ldots, i_{n+1} \in I \end{aligned}$$

X_t hay $X(t)$ được gọi là trạng thái của xích Markov tại thời điểm $t, t \in T$: tập thời gian. Nếu $X(t)$ có tính Markov và không gian trạng thái I đánh số được (đếm được) thì $X(t)$ gọi là xích Markov. Thêm vào đó, nếu $t \in T = \{1, 2, \ldots\}$ thì ta có khái niệm xích Markov với thời gian rời rạc, còn nếu $t \in T = [0, \infty)$ thì ta có khái niệm xích Markov với thời gian liên tục.

Giả sử $(X_n)_{n\geq 0}$ là xích Markov rời rạc và thuần nhất với không gian trạng thái I. Khi đó tính Markov rời rạc và thuần nhất của (X_n) có nghĩa là:

$$\begin{aligned} p_{ij} &= P(X_{n+1} = j \mid X_n = i) \\ &= P(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \end{aligned}$$

không phụ thuộc vào n.

Ta gọi $p_{ij} = P(X_{n+1} = j \mid X_n = i)$ là xác suất chuyển trạng thái sau một bước. Nghĩa là p_{ij} là xác suất có điều kiện để hệ tại thời điểm n (hiện tại) ở trạng thái i chuyển sang trạng thái j tại thời điểm $n+1$ (tương lai).

Định nghĩa 2.4: *(Quá trình Markov). Quá trình ngẫu nhiên $(X_t)_{t\geq 0}$ được gọi là quá trình ngẫu nhiên Markov nếu $\forall t_0 < t_1 < \ldots < t_{n-1} < t_n < \ldots \in T$ thì $X_n = X(t_n), n = 0, 1, 2, \ldots$ là xích Markov. Tức thỏa mãn:*

$$\begin{aligned} P(X(t_{n+1}) = j \mid X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, \ldots, X(t_0) = i_0) \\ = P(X(t_{n+1}) = j \mid X(t_n) = i_n) \end{aligned}$$

Xác suất chuyển được kí hiệu là $p(s, i, t, j)$ là xác suất để hệ tại thời điểm s (hiện tại) ở trạng thái i chuyển sang trạng thái j ở thời điểm t (tương lai) $(t > s)$ và được tính bởi công thức:

$$p(s, i, t, j) = P(X(t) = j \mid X(s) = i), t > s$$

Phân phối hữu hạn chiều của quá trình Markov $(X_t)_{t\geq 0}$ được tính bởi công thức:

$$\begin{aligned} P(X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1, X(t_0) = i_0) \\ = p(t_0, i_0).p(t_0, i_0, t_1, i_1) \ldots p(t_{n-1}, i_{n-1}, t_n, i_n) \end{aligned}$$

Định nghĩa 2.5: *(Quá trình Markov thuần nhất). Nếu xác suất chuyển chỉ phụ thuộc khoảng thời gian từ $s \to t$,tức là:*

$$p(s, i, t, j) = p(s + h, i, t + h, j)$$

thì quá trình Markov được gọi là thuần nhất.

2.2 Mô hình Markov

Xét một chuỗi các biến trạng thái quan sát $\{q_1, q_2, \ldots, q_n\}$. Một mô hình Markov sẽ chứa giả thuyết Markov về các xác suất xảy ra của các trạng thái trong chuỗi: rằng khi dự đoán một trạng thái trong tương lai, ta chỉ cần quan tâm tới trạng thái hiện tại mà những trạng thái trong quá khứ không ảnh hưởng.

$$P(q_n \mid q_{n-1}, \ldots, q_2, q_1) = P(q_n \mid q_{n-1}) \tag{2.1}$$

Phương trình 2.1 được gọi là *Giả thuyết Markov bậc 1.* Xác suất của một quan sát q_n tại thời điểm n chỉ phụ thuộc vào quan sát q_{n-1} ở thời điểm $n-1$. Một giả thuyết Markov bậc 2 là xác suất của một quan sát q_n tại thời điểm n chỉ phụ thuộc vào quan sát q_{n-1} ở thời điểm $n-1$ và quan sát q_{n-2} ở thời điểm $n-2$.

Nói chung khi nhắc đến giả thuyết Markov người ta thường ngụ ý là giả thuyết Markov bậc 1, vì trong q_{n-1} đã chứa thông tin của $q_{n-2}, q_{n-3}, \ldots$(theo hệ thức truy hồi).

Định nghĩa 2.6: *(Mô hình Markov bậc 1). Một hệ thống mà thỏa mãn giả thuyết Markov bậc 1 thì được gọi là mô hình Markov (bậc 1) và chuỗi trạng thái quan sát $\{q_i\}$ của hệ được gọi*

là chuỗi Markov (bậc 1).

Xác suất của một chuỗi trạng thái quan sát $\{q_1, q_2, \ldots, q_n\}$ dùng giả thuyết Markov có thể được biểu diễn như sau:

$$P(q_1, q_2, \ldots, q_n) = \prod_{i=1}^{n} P(q_i \mid q_{i-1}) \tag{2.2}$$

Chương 3

Mô hình Markov ẩn

Chuỗi Markov rất hữu ích khi chúng ta cần tính xác suất cho một chuỗi các sự kiện có thể quan sát được. Tuy nhiên, trong nhiều trường hợp, các sự kiện chúng ta quan tâm lại bị ẩn đi, nói cách khác, chúng ta không thể quan sát chúng một cách trực tiếp.

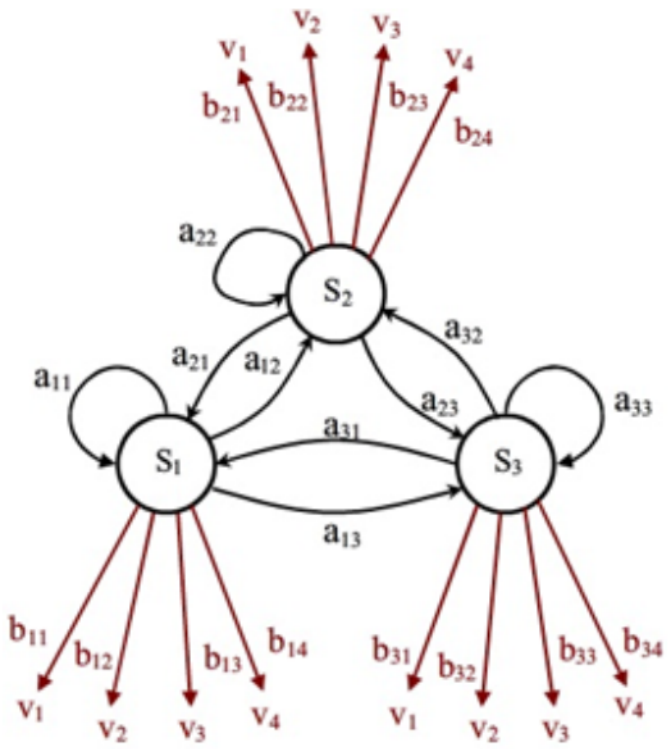
Một mô hình Markov ẩn cho phép ta tìm hiểu về cả các sự kiện được quan sát và các sự kiện ẩn mà đó chính là các yếu tố quan trọng trong một mô hình ngẫu nhiên để áp dụng tìm hiểu và nghiên cứu các mô hình trong thực tế.

3.1 Mô hình Markov ẩn

3.1.1 Khái niệm

Mô hình Markov ẩn (Hidden Markov Model - HMM) là kết quả mở rộng của khái niệm Mô hình Markov rời rạc và thuần nhất bằng cách mỗi trạng thái được gắn với một hàm phát xạ quan sát (observation distribution). Ngoài quá trình ngẫu nhiên chuyển giữa các trạng thái, tại mỗi trạng thái còn có một quá trình ngẫu nhiên nữa đó là quá trình ngẫu nhiên sinh ra một quan sát. Như vậy trong mô hình Markov ẩn có một quá trình ngẫu nhiên kép, trong đó có một quá trình không quan sát được. Tập các quan sát $O = \{O_1, O_2, ...\}$ được sinh ra bởi dãy các trạng thái $s_1, s_2, ..., s_N$ của mô hình, mà dãy các trạng thái này không thấy được, đó chính là lý do nó được gọi là mô hình Markov ẩn.

Ví dụ một mô hình Markov ẩn 3 trạng thái với các sự kiện có thể quan sát được trong mỗi trạng thái là $V = \{v_1, v_2, v_3, v_4\}$ (Hình 3.1). Khả năng (xác suất) quan sát được sự kiện v_k trong trạng thái S_j phụ thuộc vào hàm xác suất $b_j(k)$. Hàm b được gọi là hàm mật độ xác suất của các sự kiện được quan sát.
Sau đây ta sẽ xét một ví dụ đơn giản về mô hình Markov ẩn.



Hình 3.1: Mô hình Markov ẩn 3 trạng thái

3.1.2 Ví dụ

Có hai người bạn A và B liên lạc với nhau qua tin nhắn điện thoại, A ở Việt Nam còn B ở Mỹ. B là người có tâm trạng dễ thay đổi theo thời tiết. Thường B sẽ kể cho A biết tâm trạng của mình vui hay buồn qua các tin nhắn, nhưng không cho biết về tình hình thời tiết tại nơi B ở. Vậy A có thể đoán được trạng thái thời tiết ở Mỹ nếu chỉ biết xác suất thay đổi thời tiết của 2 ngày liên tiếp tại Mỹ và các khả năng để B cảm thấy vui hay buồn dựa trên thời tiết tại đó được không?
Giả sử xác suất chuyển trạng thái của thời tiết (nắng, mưa, râm) tại Mỹ được cho trong Bảng 3.1 và giả sử B cho biết sự thay đổi tâm trạng của mình (vui, buồn) dựa trên thời tiết được cho trong bảng sau:

	Vui	Buồn
Nắng	0.9	0.1
Mưa	0.2	0.8
Râm	0.6	0.4

Bảng 3.1: Xác suất thay đổi tâm trạng theo thời tiết của B

Ở đây thời tiết ở Mỹ là chi tiết bị ẩn đối với A (người bạn ở Việt Nam). Chi tiết mà A biết được (có thể "quan sát"được) đó là các tin nhắn nói về tâm trạng của mình từ B. Việc cần làm là tìm xác suất của thời tiết $q_i \in N, M, R$ chỉ có thể dựa trên các quan sát về tâm trạng của B, ta kí hiệu là O_i , với $O_i = V$ nếu B cảm thấy vui với thời tiết ở ngày i và $O_i = B$ nếu B cảm thấy buồn với thời tiết ở ngày i. Ta có xác suất có điều kiện $P(q_i \mid O_i)$ có thể được viết theo

luật Bayes như sau:

$$P(q_i \mid O_i) = \frac{P(O_i \mid q_i)P(q_i)}{P(O_i)} \tag{3.1}$$

Câu hỏi
a. Giả sử ngày đầu tiên hai người nhắn tin thì thời tiết ở Mỹ là trời nắng. Ngày tiếp theo B nhắn cho A rằng mình cảm thấy vui với thời tiết hiện tại. Biết rằng xác suất trước đó (trước khi hai người bắt đầu nhắn tin) để B thấy vui là 0.5. Tính xác suất để ngày thứ hai hai người nhắn tin là trời mưa?

$$\begin{aligned}
P(q_2 = M \mid q_1 = N, O_2 = V) &= \frac{P(q_2 = M, q_1 = N | O_2 = V)}{P(q_1 = N | O_2 = V)} \\
&= \frac{P(q_2 = M, q_1 = N \mid O_2 = V)}{P(q1 = N)} && (q_1, O_2 \text{ độc lập}) \\
&= \frac{P(O_2 = V \mid q_1 = N, q_2 = M)P(q_2 = M, q_1 = N)}{P(q_1 = N | O_2 = V)} \\
& && \text{(Luật Bayes)} \\
&= \frac{P(O_2 = V \mid q_2 = M)P(q_2 = M, q_1 = N)}{P(q_1 = N)P(O_2 = V)} && \text{(Giả thiết Markov)} \\
&= \frac{P(O_2 = V \mid q_2 = M)P(q_2 = M \mid q_1 = N)P(q_1 = N)}{P(q_1 = N)P(O_2 = V)} \\
&= \frac{P(O_2 = V \mid q_2 = M)P(q_2 = M \mid q_1 = N)}{P(O_2 = V)} \\
&= \frac{0.2 \times 0.05}{0.5} = 0.02
\end{aligned}$$

b. Không đề cập tới xác suất trước đó của tâm trạng của B. Hãy tính các khả năng thởi tiết xảy ra vào ngày hôm sau (các giả thiết khác cho như câu a)

Trước hết ta có phân tích: với n ngày nhắn tin và chuỗi thời tiết ở Mỹ là $q = \{q_1, q_2, ..., q_n\}$, chuỗi diễn biến tâm trạng của B là $O = \{O_1, O_2, ..., O_n\}$ thì ta có:

$$P(q_1, \ldots, q_n \mid O_1, \ldots, O_n) = \frac{P(O_1, \ldots, O_N \mid q_1, \ldots, q_n)P(q_1, \ldots, q_n)}{P(O_1, \ldots, O_n)} \tag{3.2}$$

Sử dụng xác suất $P(q_1, \ldots, q_n)$ của chuỗi thời tiết Markov và xác suất $P(O_1, \ldots, O_n)$ của quan sát thực tế các thay đổi tâm trạng để tính.
Nếu giả sử rằng O_i và q_j là độc lập với mọi $i \neq j$ thì

$$P(O_1, \ldots, O_N \mid q_1, \ldots, q_n) = \prod_{i=1}^{n} P(O_i \mid q_i) \tag{3.3}$$

Để đưa ra kết luận cho trạng thái thời tiết ở Mỹ dựa trên những cập nhật về sự thay đổi tâm trạng của B về thời tiết bên ngoài, ta giả sử rằng xác suất chuỗi quan sát tâm trạng của B $P(O_1, \ldots, O_N)$ là độc lập với chuỗi thời tiết chúng ta dự đoán $P(q_1, \ldots, q_N)$. Định nghĩa khả năng xảy ra L như sau

$$L(q_1, ..., q_n \mid O_1, ..., O_n) = P(O_1, ..., O_n \mid q_1, ..., q_n)P(q_1, ..., q_n) \tag{3.4}$$

Với giả thuyết Markov bậc 1 biểu thức trên trở thành

$$L(q_1, ..., q_n \mid O_1, ..., O_n) = \prod_{i=1}^{n} P(O_i \mid q_i) \prod_{i=1}^{n} P(q_i \mid q_{i-1}) \tag{3.5}$$

Từ công thức 3.5, ta có thể áp dụng để trả lời yêu cầu của câu b) như sau:

- Khả năng nắng của thời tiết ngày tiếp theo là

$$L(q_2 = N|q_1 = N, O_2 = V) = P(O_2 = V|q_2 = N)P(q_2 = N|q_1 = N)$$
$$= 0.9 \times 0.8 = 0.72$$

- Khả năng mưa của thời tiết ngày tiếp theo là

$$L(q_2 = M|q_1 = N, O_2 = V) = P(O_2 = V|q_2 = M)P(q_2 = M|q_1 = N)$$
$$= 0.2 \times 0.05 = 0.01$$

- Khả năng râm của thời tiết ngày tiếp theo là

$$L(q_2 = R|q_1 = N, O_2 = V) = P(O_2 = V|q_2 = R)P(q_2 = R|q_1 = N)$$
$$= 0.6 \times 0.15 = 0.09$$

Như vậy, với điều kiện là ngày đầu tiên trời nắng và ở ngày thứ hai B thấy vui thì xác suất cao nhất của thời tiết ngày hôm sau đó là trời nắng.

c. Giả sử ba ngày liên tiếp (tính từ ngày đầu tiên hai người nhắn tin) B đều nhắn cho A rằng cảm thấy buồn. Tính khả năng xảy ra để thời tiết cho 3 ngày đó lần lượt là $(q_1 = N, q_2 = R, q_3 = N)$ với giả thiết xác suất của 3 trạng thái thời tiết (N, M, R) ở ngày đầu tiên là tương đương nhau, tức

$$P(q_1 = N|q_0) = P(q_1 = N) = 1/3,$$
$$P(q_1 = M|q_0) = P(q_1 = M) = 1/3,$$
$$P(q_1 = R|q_0) = P(q_1 = R) = 1/3.$$

Khi đó khả năng để thời tiết cho 3 ngày lần lượt là $(q_1 = N, q_2 = R, q_3 = N)$ như sau

$$\begin{aligned} L(q_1 = N, q_2 = R, q_3 = N|O_1 = B, O_2 = B, O_3 = B) \\ = P(O_1 = B|q_1 = N)P(O_2 = B|q_2 = R)P(O_3 = B|q_3 = N) \\ * P(q_1 = N|q_0)P(q_2 = R|q_1 = N)P(q_3 = N|q_2 = R) \\ = 0.1 \times 0.4 \times 0.1 \times 1/3 \times 0.15 \times 0.2 = 4.10^{-5} \end{aligned}$$

3.2 Các thành phần và hoạt động của mô hình Markov ẩn

3.2.1 Các thành phần của mô hình Markov ẩn

Một mô hình Markov ẩn được đặc trưng bởi 5 thành phần cơ bản sau

i) N - số trạng thái trong mô hình. $S = \{s_1, ..., s_N\}$ là tập các trạng thái.

ii) M - số ký hiệu quan sát có thể. $V = \{v_1, ..., v_M\}$ là tập các ký hiệu quan sát có thể.

iii) $A = \{a_{ij}\}$ - xác suất chuyển trạng thái

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$$

$$\begin{cases} \sum_{j=1}^{N} a_{ij} = 1 & i = \overline{1, N} \\ a_{ij} \geq 0 & i, j = \overline{1, N} \end{cases}$$

iv) $B = b_j(k)$ - xác suất phát xạ quan sát trong mỗi trạng thái. $b_j(k)$ là xác suất của quan sát v_k tại trạng thái s_j ở thời điểm t

$$b_j(k) = P(v_k) \text{ tại thời điểm } t | q_t = s_j$$

$$\begin{cases} \sum_{k=1}^{M} b_j(k) = 1 & j = \overline{1, N} \\ b_j(k) \geq 0 & i, j = \overline{1, N}; k = \overline{1, M} \end{cases}$$

v) $\pi = \{\pi_1, ..., \pi_N\}$ là tập các phân bố xác suất cho trạng thái khởi đầu, π_i là xác suất để trạng thái s_i được chọn tại thời điểm khởi đầu $t = 1$

$$\pi_i = P(q_1 = s_i)$$

$$\begin{cases} \sum_{i=1}^{N} \pi_i = 1 & i = \overline{1, N} \\ \pi_i \geq 0 & i, i = \overline{1, N} \end{cases}$$

vi) $O = \{O_1, O_2, \ldots, O_R\}$ là một dãy quan sát với O_t là một ký hiệu quan sát trong tập V và R là số lượng quan sát trong dãy. Mô hình Markov ẩn như vậy được ký hiệu là $\lambda = (A, B, \pi)$. Với A, B, π là các tham số.

Trong ví dụ về người dự đoán thời tiết ở nơi bạn mình sống dựa trên các tin nhắn thể hiện tâm trạng của bạn, ta có các dữ kiện sau:

- $N = 3, S = \{s_1 = \text{Nắng}, s_2 = \text{Mưa}, s_3 = \text{Râm }\}$
- $M = 2, V = \{v_1 = \text{vui } (V), v2 = \text{buồn } (B)\ \}$
- A là các xác suất chuyển trạng thái của thời tiết qua các ngày.
- $B = \{b_j(k)\}$ là xác suất quan sát được tâm trạng của B là v_k khi thời tiết là s_j tại thời điểm t.

3.2.2 Hoạt động của mô hình Markov ẩn

Hoạt động của một mô hình Markov ẩn được mô tả như sau

1. Chọn một trạng thái khởi đầu $q_1 = s_i$ tương ứng với xác suất cho trạng thái khởi đầu π
2. Gán $t = 1$.
3. Chọn $O_t = v_k$ tương ứng với xác suất quan sát tại trạng thái s_i , tức là $b_i(k)$.
4. Chuyển sang trạng thái mới $q_{t+1} = s_j$ tương ứng với xác suất chuyển trạng thái a_{ij} .
5. Gán $t = t + 1$ và quay lại bước 3 nếu $t < T$. Nếu ngược lại thì kết thúc.

Bộ ba $\lambda = (A, B, \pi)$ được coi là bộ ký pháp gọn để biểu diễn một mô hình Markov ẩn. A, B, π được gọi là *các tham số* của mô hình λ.

3.2.3 Các công thức được sử dụng trong mô hình Markov ẩn

- Xác suất của một dãy trạng thái: Xác xuất của một dãy trạng thái $q = \{q_1, ..., q_N\}$ từ mô hình HMM với các tham số λ tương ứng là tích của các xác suất chuyển tiếp của mỗi trạng thái cho bởi công thức sau

$$P(q|\lambda) = \pi_{q_1} \prod_{n-1}^{N-1} a_{q_n q_{n+1}} = \pi_{q_1} a_{q_1 q_2} \dots a_{q_{N-1} q_N}$$

- Khả năng của một dãy quan sát sinh bởi dãy trạng thái tương ứng: Cho dãy quan sát $O = \{O_1, ..., O_N\}$ và dãy các trạng thái $q = \{q_1, ...q_N\}$ được xác định bởi mô hình HMM với các tham số λ khả năng của dãy O với điều kiện q được cho bởi

$$P(O|q, \lambda) = \prod_{n=1}^{N} P(O_n|q_n, \lambda) = b_{q_1 O_1} \dots b_{q_N O_N}$$

  là một tích các xác suất xạ quan sát.

- Khả năng chung của một dãy quan sát O và dãy trạng thái q tương ứng đó là xác suất để O và q xảy ra đồng thời

$$P(O, q|\lambda) = P(O|q, \lambda)P(q|\lambda) \qquad \text{(Bayes)}$$

- Khả năng của một dãy quan sát $O = \{O_1, ..., O_N\}$ có mối liên hệ với một HMM với các tham số λ xác định bởi

$$P(O|\lambda) = \sum_{\forall q} P(O, q|\lambda)$$

  đó là tổng của các khả năng chung của dãy tất cả các dãy trạng thái có thể của q cho bởi mô hình.

3.3 Ba bài toán cơ bản của mô hình Markov ẩn

3.3.1 Bài toán 1: Bài toán ước lượng (Likelihood Computation)

Bài toán: Cho mô hình Markov ẩn $\lambda = (A, B, \pi)$ và một dãy quan sát $O = \{O_1, O_2...\}$, cần tính xác suất $P(O|\lambda)$.

Giả sử ta có dãy quan sát $O = \{O_1, O_2, ..., O_T\}$ độ dài T và dãy các trạng thái tương ứng của mô hình Markov ẩn $q = \{q_1, q_2, ..., q_T\}$. Khi đó, xác suất để dãy quan sát O được sinh ra bởi mô hình λ là:

$$\begin{aligned} P(O|\lambda) &= \sum_{\forall q} P(O, q|\lambda) \\ &= \sum_{\forall q} P(O|q, \lambda).P(q|\lambda) \\ &= \sum_{\forall q} \left[\prod_{t=1}^{T} P(O_t|q_t, \lambda).\lambda_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \right] \end{aligned}$$

Ta có thể tính xác suất này bằng cách liệt kê tất cả các khả năng có thể của dãy trạng thái q . Tuy nhiên cách này có độ phức tạp tính toán rất lớn.
Thật vậy, từ $q_t \in \{s_1, ..., s_N\}$ suy ra có N cách chọn mỗi q_t , do đó có N^T cách chọn dãy $q = \{q_1, ..., q_T\}$. Mặt khác, theo 3.7 có $2T - 1$ phép nhân với mỗi cách chọn dãy q . Vậy độ phức tạp tính toán sẽ là $\approx O(2T.N^T)$. Vấn đề này sẽ được khắc phục trong thuật toán tiến - lùi sau đây.

- **Thuật toán tiến**
  Gọi biến tiến $\alpha_t(i) = P(O_1, O_2, ..., O_t, q_t = s_i|\lambda)$ là xác suất biến quan sát O tới thời điểm $t : O = O_1, ..., O_t$, tại trạng thái s_i được sinh bởi mô hình λ.
  Các giá trị $\alpha_t(i)$ được tính bằng thuật toán đệ quy sau

  B1. Khởi tạo

  $$\alpha_1(i) = \pi_i b_i(O_1), 1 \leq i \leq N \tag{3.6}$$

  B2. Tính các $\alpha_{t+1}(j)$ bằng phương pháp đệ quy

  $$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \lambda_t(i) a_{ij} \right] b_j(O_{t+1}) 1 \leq t \leq T-1; 1 \leq j \leq N \tag{3.7}$$

  B3. Kết thúc

  $$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{3.8}$$

  Thuật toán dừng lại ở $B3$ khi $t = T$.
  Thuật toán có $(T-1)(N+1)N$ phép nhân. Vậy độ phức tạp tính toán của cách tính theo các biến tiến $\alpha_t(i)$ là $\approx O(N^2.T)$ thấp hơn so với độ phức tạp $O(2T.N^T)$

- **Thuật toán lùi**
  Gọi biến lùi $\beta_t(i) = P(O_{t+1}, O_{t+2}, ..., O_T | q_t = s_i, \lambda)$ là xác suất cảu dãy quan sát O từ thời điểm $t+1$ đến thời điểm $T : O = O_{t+1}, O_{t+2}, ..., O_T$ với điều kiện là mô hình ở trạng thái s_i tại thời điểm t. Các $\beta_t(i)$ được tính bằng thuật toán đệ quy sau

  B1. Khởi tạo

$$\beta_T(i) = 1, 1 \leq i \leq N \tag{3.9}$$

  B2. Tính các $\beta_t(j)$ bằng phương pháp đệ quy

$$\beta_t(j) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{3.10}$$
$$t = T-1, T-2, \ldots, 1; 1 \leq i \leq N$$

  B3. Kết thúc

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i) \tag{3.11}$$

  Thuật toán tính theo biến lùi $\beta_t(j)$ cũng có độ phức tạp tính toán $\approx O(N^2.T)$.
  Bằng thuật toán tiến-lùi, ta có thể tính xác suất

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i). \tag{3.12}$$

3.3.2 Bài toán 2: Bài toán giải mã (Decoding Computation)

Bài toán: Với dãy quan sát $O = \{O_1, O_2, ...\}$ và mô hình Markov ẩn $\lambda = (A, B, \pi)$. Làm thể nào để có thể tìm được dãy trạng thái tương ứng $q = \{q_1, q_2, ..., \}$ tối ưu nhất theo một tiêu chuẩn nào đó?

Một phương pháp thông dụng được dùng để giải quyết bài toán này là dùng thuật toán tìm kiếm Viterbi để tìm ra một dãy các trạng thái tối ưu duy nhất.
Đặt $\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P(q_1 q_2 ... q_{t-1} q_t = s_i, O_1 O_2 ... O_t | \lambda)$ là biến cố xác suất cao nhất tại mọi thời điểm t tương ứng với dãy trạng thái $q_1, q_2, ..., q_{t-1}$ kết thúc tại trạng thái $q_i = s_i$. Các biến $\delta_{t+1}(j)$ được tính bằng phương pháp đệ quy dựa trên các tính toán trước đó

$$\delta_{t+1}(j) = \left[\max_{a \leq i \leq N} \delta_t(i) a_{ij} \right] b_j(O_{t+1}) \tag{3.13}$$

Để lưu vết các trạng thái của dãy các trạng thái tối ưu ta dùng mảng $\psi_t(j)$, khi thuật toán kết thúc các phần tử trong mảng chính là các trạng thái của dãy q cần tìm. Thuật toán tìm kiếm Viterbi mô tả như sau

B1. Khởi tạo

$$\delta_1(i) = \pi_i b_i(O_i), 1 \leq i \leq N \tag{3.14}$$

B2. Đệ quy

$$\delta_t(j) = \left[\max_{1\le i\le N} \delta_{t-1}(i)a_{ij}\right] b_j(O_t), \qquad 2 \le t \le T; 1 \le j \le N \tag{3.15}$$

$$\psi_t(j) = \arg\max_{1\le i\le N} \left[\delta_{t-1}(i)a_{ij}\right], \qquad 2 \le t \le T; 1 \le j \le N \tag{3.16}$$

B3. Kết thúc

$$P^*(O|\lambda) = \max_{1\le i\le N} [\delta_T(i)]$$

B4. Truy hồi các trạng thái

$$q_t^* = \psi_{t+1}(q_{t+1}^*, \qquad t = T-1, T-2, ..., 1 \tag{3.17}$$

Kết thúc thuật toán, các q_t^* chính là các trạng thái của dãy cần tìm. Theo 3.15 để tính tất cả $\delta_t(j)$ cần $(N+1)N(T-1)$ phép nhân, để tính tất cả $\psi_t(j)$ cần $(T-1)N^2$ phép nhân. Vậy độ phức tạp của thuật toán Viterbi $\approx O(2T.N^2.)$.

Ví dụ bài toán 2. Xét mô hình HMM về thời tiết đã nói trong phần trước, nhưng không biết về thời tiết cho đến khi quan sát được. Trong ba ngày đầu tiên, quan sát các thay đổi tâm trạng của B là { buồn, vui, vui } ($\{B, V, V\}$). Sử dụng thuật toán Viterbi, tìm chuỗi thời tiết có khả năng nhất ứng với chuỗi quan sát tâm trạng thu được. (Giả thiết rằng khả năng của ba trạng thái thời tiết ở ngày đầu tiên là tương đương)

B1. Khởi đầu: $n = 1$

$$\delta_1(N) = \pi_N.b_N(B) = \frac{1}{3} \times 0.1 = 0.033$$
$$\psi_1(N) = 0$$
$$\delta_1(M) = \pi_M.b_M(B) = \frac{1}{3} \times 0.8 = 0.267$$
$$\psi_1(M) = 0$$
$$\delta_1(R) = \pi_R.b_R(B) = \frac{1}{3} \times 0.a = 0.133$$
$$\psi_1(R) = 0$$

B2. Đệ quy: Tính toán khả năng của trạng thái "N"từ tất cả các trường hợp có thể của ba trạng thái trước đó và chọn ra một trạng thái có khả năng cao nhất để tiếp tục, ta có:

+ Với $n = 2$

$$\begin{aligned}\delta_2(N) &= \max\{\delta_1(N).a_{N,N}; \delta_1(M).a_{M,N}; \delta_1(R).a_{R,N}\}b_N(V) \\ &= \max\{0.033 \times 0.8; 0.267 \times 0.2; 0.133 \times 0.2\} \times 0.9 \\ &= 0.0534 \times 0.9 = 0.04806\end{aligned}$$
$$\psi_2(N) = N$$

Thực hiện tương tự đối với các trạng thái M và R, ta có:

$$\begin{aligned}
\delta_2(M) &= \max\{\delta_1(N).a_{N,M}; \delta_1(M).a_{M,M}; \delta_1(R).a_{R,M}\}b_M(V) \\
&= \max\{0.033 \times 0.05; 0.267 \times 0.6; 0.133 \times 0.3\} \times 0.2 \\
&= 0.1602 \times 0.2 = 0.03204 \\
\psi_2(M) &= M \\
\delta_2(R) &= \max\{\delta_1(N).a_{N,R}; \delta_1(M).a_{M,R}; \delta_1(R).a_{R,R}\}b_R(V) \\
&= \max\{0.033 \times 0.15; 0.267 \times 0.2; 0.133 \times 0.5\} \times 0.6 \\
&= 0.0665 \times 0.6 = 0.0399 \\
\psi_2(R) &= R
\end{aligned}$$

+ $n = 3$

$$\begin{aligned}
\delta_3(N) &= \max\{\delta_2(N).a_{N,N}; \delta_2(M).a_{M,N}; \delta_2(R).a_{R,N}\}b_N(V) \\
&= \max\{0.04806 \times 0.8; 0.03204 \times 0.2; 0.0399 \times 0.2\} \times 0.9 \\
&= 0.03845 \times 0.9 = 0.03461 \\
\psi_3(N) &= N \\
\delta_3(M) &= \max\{\delta_2(N).a_{N,M}; \delta_2(M).a_{M,M}; \delta_2(R).a_{R,M}\}b_M(V) \\
&= \max\{0.04806 \times 0.05; 0.03204 \times 0.6; 0.0399 \times 0.3\} \times 0.2 \\
&= 0.01922 \times 0.2 = 0.0038 \\
\psi_3(M) &= M \\
\delta_3(R) &= \max\{\delta_2(N).a_{N,R}; \delta_2(M).a_{M,R}; \delta_2(R).a_{R,R}\}b_R(V) \\
&= \max\{0.04806 \times 0.15; 0.03204 \times 0.2; 0.0399 \times 0.5\} \times 0.6 \\
&= 0.02 \times 0.6 = 0.0012 \\
\psi_3(R) &= R
\end{aligned}$$

Sau cùng, ta thu đường một chuỗi chu trình có khả năng nhất trong mỗi trạng thái của mô hình.

B3. Kết thúc
Toàn bộ chu trình có khả năng nhất đã được xác định, ta bắt đầu tìm kiếm trạng thái cuối cùng của dãy có khả năng nhất:

$$\begin{aligned}
P^*(O|\lambda) &= \max\left[\delta_3(i)\right] = \delta_3(N) = 0.03461 \\
q_3^* &= \arg\max\left[\delta_3(i)\right] = N
\end{aligned}$$

B4. Truy hồi các trạng thái
Dãy các trạng thái tốt nhất có thể thu được từ hàm ψ là:
+ Với $n = N - 1 = 2$:

$$q_2^* = \psi_3(q_3^*) = \psi_3(N) = N$$

+ Với $n = N - 1 = 1$:

$$q_1^* = \psi_2(q_2^*) = \psi_2(N) = M$$

Như vậy dãy thời tiết có khả năng nhất là:

$$q^* = \{q_1^*, q_2^*, q_3^*\}$$

3.3.3 Bài toán 3: Bài toán huấn luyện (Training HMM)

Bài toán: Bài toán tối ưu các tham số của mô hình. Tìm cách nào để điều chỉnh các tham số A, B, π để được xác suất $P(O|\lambda)$ lớn nhất?

Đây là bài toán khó nhất của mô hình Markov ẩn. Giải pháp cho bài toán này là thủ tục huấn luyện lặp Baum-Welch. Chọn biến $\gamma_t(i) = P(q_t = s_i|O, \lambda)$ là xác suất để mô hình ở trạng thái s_i tại thời điểm t với dãy quan sát O và mô hình λ đã cho. Với định nghĩa trên, biến $\gamma_t(i)$ được biểu diễn thông qua hai biến tiến và lùi như sau

$$\gamma_t(i) = \frac{P(q_t = s_i|O, \lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{3.18}$$

Từ công thức 3.18 rút ra được $\sum \gamma_t(i) = 1$.
Với $\gamma_t(i)$ có thể tìm được tại thời điểm t xác suất lớn nhất của dãy $\{O_1, O_2, ..., O_t\}$ là

$$q_t = \arg\max\left[\gamma_t(i)\right], \qquad 1 \le i \le N; 1 \le t \le T \tag{3.19}$$

Chọn biến $\xi_t(i, j)$ là xác suất mô hình ở trạng thái s_i tại thời điểm t và ở trạng thái s_j tại thời điểm $t + 1$ với mô hình λ và dãy quan sát O cho trước, tức là

$$\xi_t(i, j) = P(q_t = s_i, q_{t+1} = j|O, \lambda) \tag{3.20}$$

Với biến tiến $\alpha_t(i)$ và biến lùi $\beta_t(j)$ được định nghĩa như trên, $\xi_t(i, j)$ có thể biểu diễn như sau

$$\begin{aligned}\xi_t(i, j) &= \frac{P(q_t = s_i, q_{t+1} = j|O, \lambda}{P(O|\lambda} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1}\beta_{t+1}(j)}{P(O|\lambda} \\ &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1}\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)a_{ij}b_j(O_{t+1}\beta_{t+1}(j)}\end{aligned} \tag{3.21}$$

Từ định nghĩa của $\gamma_t(i)$ ta có

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \tag{3.22}$$

Từ các công thức trên có thể nhận thấy:
$\sum_{t=1}^{T-1} \xi_t(i, j)$ = khả năng để mô hình chuyển trạng thái từ s_i sang s_j.
$\sum_{t=1}^{T-1} \gamma_t(i)$ = khả năng để mô hình chuyển trạng thái từ s_i.

Từ các công thức trên ta có tập các công thức dùng để điều chỉnh (re-estimation) các tham số của mô hình Markov ẩn như sau:

$\overline{\pi_i}$ = khả năng mô hình ở trạng thái s_i tại thời điểm $(t = 1)$

$$\overline{\pi_i} = \gamma_1(i) \tag{3.23}$$

$\overline{a_{ij}}$= (khả năng chuyển trang thái từ s_i sang s_j)/(khả năng chuyển từ trạng thái s_i)

$$\Rightarrow \overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.24}$$

$\overline{b_j(v_k)}$ = (khả năng ở tại trạng thái s_i với ký hiệu quan sát v_k)/(khả năng ở tại trạng thái v_i)

$$\Rightarrow \overline{b_j(v_k)} = \frac{\sum_{t=1, O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{3.25}$$

Với một mô hình $\lambda = (A, B, \pi)$ đầu tiên, sử dụng các công thức, 3.23, 3.24 và 3.25 để tính toán bộ tham số mới $\lambda = (A, B, \pi)$. Theo [7] ta đã chứng minh được rằng:

- Hoặc là mô hình khởi điểm λ được định nghĩa chính xác là mô hình hội tụ và do đó $\lambda = \overline{\lambda}$.
- Hoặc là mô hình mới có $P(O|\overline{\lambda}) > P(O|\lambda)$.

Dựa vào chứng minh này, dùng $\overline{\lambda}$ thay thể cho λ và lặp lại các tính toán (3.23), (3.24), (3.25) sẽ cải thiện được xác suất $P(O|\lambda)$ cho tới thời điểm thuật toán hội tụ.

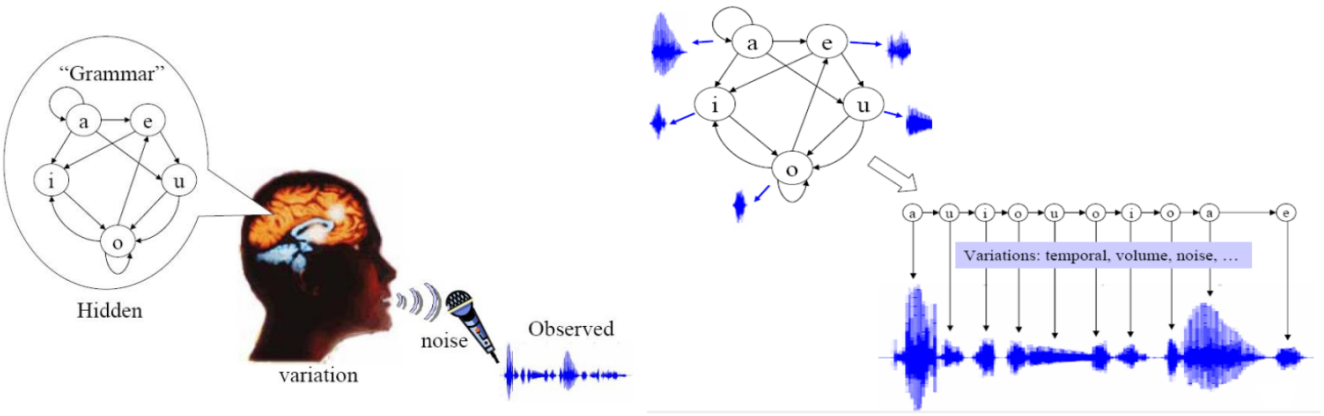
Trong quá trình tính toán, sau mỗi lần lặp các biểu thức sau đây luôn được thỏa mãn

$$\sum_{i=1}^{N} \overline{\pi_i} = 1, \sum_{j=1}^{N} \overline{a_{ij}} = 1, \sum_{k=1}^{N} \overline{b_j}(k) = 1 \tag{3.26}$$
$$1 \le i \le N, 1 \le j \le N$$

3.4 Một số ứng dụng của mô hình Markov ẩn

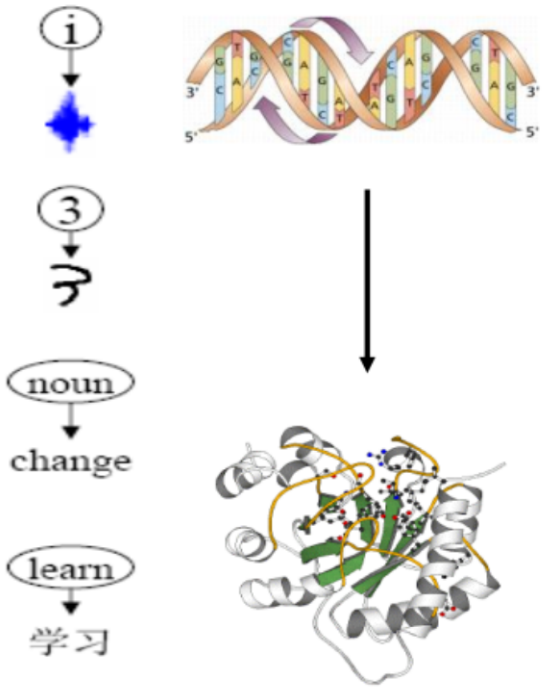
Mô hình Markov ẩn có rất nhiều ứng dụng trong thực tế:

- Các bài toán dự báo: dự báo thời tiết, dự đoán biến động thị trường chứng khoán,. . .
- Các bài toán nhận dạng: Nhận dạng tiếng nói, nhận dạng thực thể, nhận dạng chữ viết tay, . . .



Hình 3.2: Ứng dụng của HMM trong nhận dạng giọng nói

- Tin sinh học và hệ gene học: Dự đoán vùng mang mã trên một trình từ gene, xác định các họ gene hoặc họ protein liên quan, mô phỏng cấu trúc không gian của protein từ trình tự aminoacid, . . .
- Xử lí tín hiệu, phân tích dữ liệu và nhận dạng mẫu.



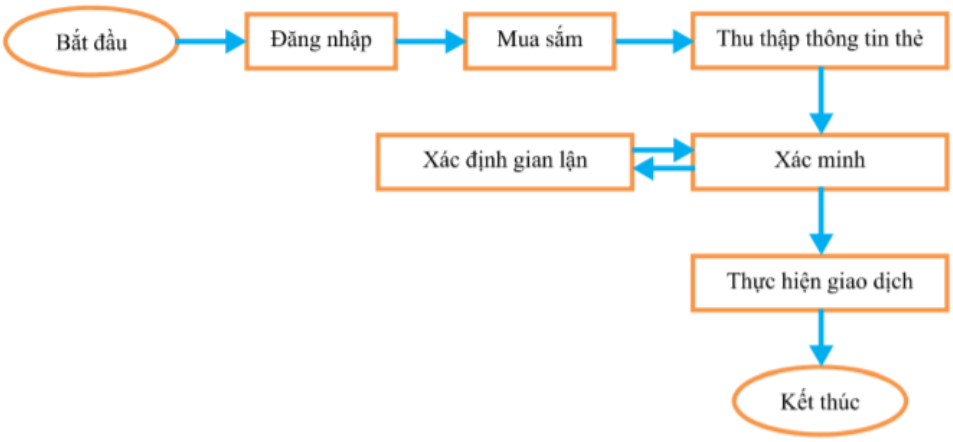
Hình 3.3: Ứng dụng của HMM trong tin sinh học

Chương 4

Phát hiện gian lận sử dụng mô hình Markov ẩn

4.1 Phát hiện gian lận sử dụng mô hình Markov ẩn

Dựa trên mô hình Markov ẩn, việc phát hiện gian lận không sử dụng xác minh thông tin cá nhân, chữ ký... mà chỉ sử dụng thói quen chi tiêu của khách hàng. Khi các đối tượng gian lận, họ thường tìm cách đánh cắp tiền sao cho số tiền lấy được là nhiều nhất và trong một thời gian ngắn trước khi bị phát hiện. Vì vậy, hệ thống sẽ sử dụng các thông tin trong lịch sử giao dịch của khách hàng trước đó như loại mặt hàng, số lượng, tần suất giao dịch, số tiền giao dịch, địa điểm mua hàng ... để so sánh với giao dịch hiện tại xem có sự bất thường nào không. Nếu có, hệ thống sẽ đưa ra cảnh báo và yêu cầu người dùng xác minh danh tính hoặc yêu cầu nhập một mã xác nhận nào đó. Nếu xác nhận thành công, hệ thống sẽ cho tiếp tục giao dịch, ngược lại, đó được coi như là một giao dịch gian lận.



Hình 4.1: Xác định gian lận sử dụng HMM[6]

Trong phần này, ta áp dụng các thuật toán của bài toán ước lượng và bài toán giải mã trình bày trong chương trước. Một mô hình Markov ẩn được ký hiệu là $\lambda = (A, B, \pi)$. Với A, B, π là các tham số đã được xác định ở chương trước.

Trong thực tế, mỗi khách hàng mua rất nhiều loại mặt hàng khác nhau và khó có thể dự đoán. Mặt khác, số tiền giao dịch của khách hàng là rõ ràng. Vì vậy ta có thể tiếp cận bài toán phát hiện gian lận thẻ tín dụng như sau:

- Phân cụm các loại mặt hàng thành N cụm. Đối với mỗi khách hàng có một cách phân cụm khác nhau. Ví dụ dựa vào lịch sử giao dịch của một khách hàng, ta phân loại các loại mặt hàng thành ba cụm là: Hàng tạp hóa; hàng điện tử và các mặt hàng còn lại thành một cụm. Ta coi mỗi cụm là một trạng thái s_i và ta có N trạng thái. Và cũng dựa vào lịch sử giao dịch, ta xây dựng được ma trận xác suất chuyển A giữa các trạng thái. Ví dụ như sau khi mua một hàng tạp hóa, xác suất khách hàng tiếp tục mua một (hàng tạp hóa; hàng điện tử; còn lại) là (0.7; 0.1;0.2).

- Đối với số tiền chi tiêu cho mỗi giao dịch, ta lại phân cụm số tiền chi tiêu thành M cụm. Ví dụ phân thành 3 cụm là (0đ – 500.000đ), (500.000đ – 3.000.000đ), và lớn hơn 3.000.000đ. Ta coi mỗi cụm là một ký hiệu quan sát v_i. Với mỗi trạng thái s_i ta có xác suất phát xạ quan sát B tương ứng với M ký hiệu quan sát. Ví dụ khi đi mua hàng tạp hóa, xác suất số tiền trả cho giao dịch tương ứng với 3 cụm trên là (0.8; 0.19; 0.01).

- Kết hợp với một phân phối ban đầu phù hợp của các trạng thái ta có mô hình Markov ẩn $\lambda = (A, B, \pi)$ tướng ứng.

Mô hình trên được xây dựng dựa trên việc giao dịch của khách hàng. Giả sử ta có một dãy quan sát số tiền của các giao dịch là $O = \{O_1,\ O_2, \ldots, O_R\}$. Xác suất của dãy quan sát đó là:

$$\alpha_1 = P(O_1, O_2, \ldots, O_R|\lambda)$$

Khi khách hàng thực hiện giao dịch tiếp theo, ta có quan sát O_{R+1}, ta loại bỏ quan sát O_1 và được dãy quan sát $O = \{O_2, O_3, \ldots, O_{R+1}\}$. Xác suất của dãy quan sát này là:

$$\alpha_2 = P\left(O_2, O_3, \ldots, O_{R+1}|\lambda\ \right)$$

Và ta ký hiệu: $\Delta\alpha = \alpha_1 - \alpha_2$
Nếu $\Delta\alpha \leq 0$ thì giao dịch có quan sát O_{R+1} được coi là giao dịch bình thường.
Nếu $\Delta\alpha > 0$ thì giao dịch có thể là một giao dịch gian lận. Giao dịch đó là gian lận nếu với một ngưỡng sai khác cho phép kí hiệu là Z và

$$\frac{\Delta\alpha}{\alpha_1} \geq Z$$

Tuy nhiên, trong thực tế các trường hợp gian lận có đặc điểm thực hiện nhiều giao dịch trong một khoảng thời gian ngắn. Vì vậy ta có thể xác định gian lận bằng nhiều giao dịch sau giao dịch có quan sát O_R như sau:
Ta ký hiệu: $\beta_2 = P\left(O_{R+1},\ O_{R+2}, \ldots, O_{2R}|\lambda\right)$ và

$$\Delta\beta = \alpha_1 - \beta_2$$

Tương tự như trên, nếu $\Delta\beta > 0$ và $\frac{\Delta\beta}{\alpha_1} \geq Z$ thì hệ thống sẽ đưa ra cảnh báo các giao dịch là gian lận.

4.2 Áp dụng phương pháp

Phương pháp gồm 3 phần chính:

- Phân cụm dữ liệu
- Đào tạo dữ liệu
- Phát hiện gian lận

Cụ thể như sau:

4.2.1 Thuật toán phân cụm

Ở phần 4.1, ta có nhắc đến bước phân cụm dữ liệu. Để hệ thống có thể cảnh báo một cách chính xác, chúng ta đào tạo dữ liệu một cách cảm tính, ngẫu nhiên thì chắc chắn hiệu quả sẽ không cao. Vì vậy, chúng ta cần có một cách phân cụm dữ liệu sao cho hiệu quả. Và ở đây, ta sẽ sử dụng thuật toán phân cụm K-mean (K-mean clustering) [8] để nâng cao hiệu quả của hệ thống. Ý tưởng thuật toán như sau:

Bước 1: Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.

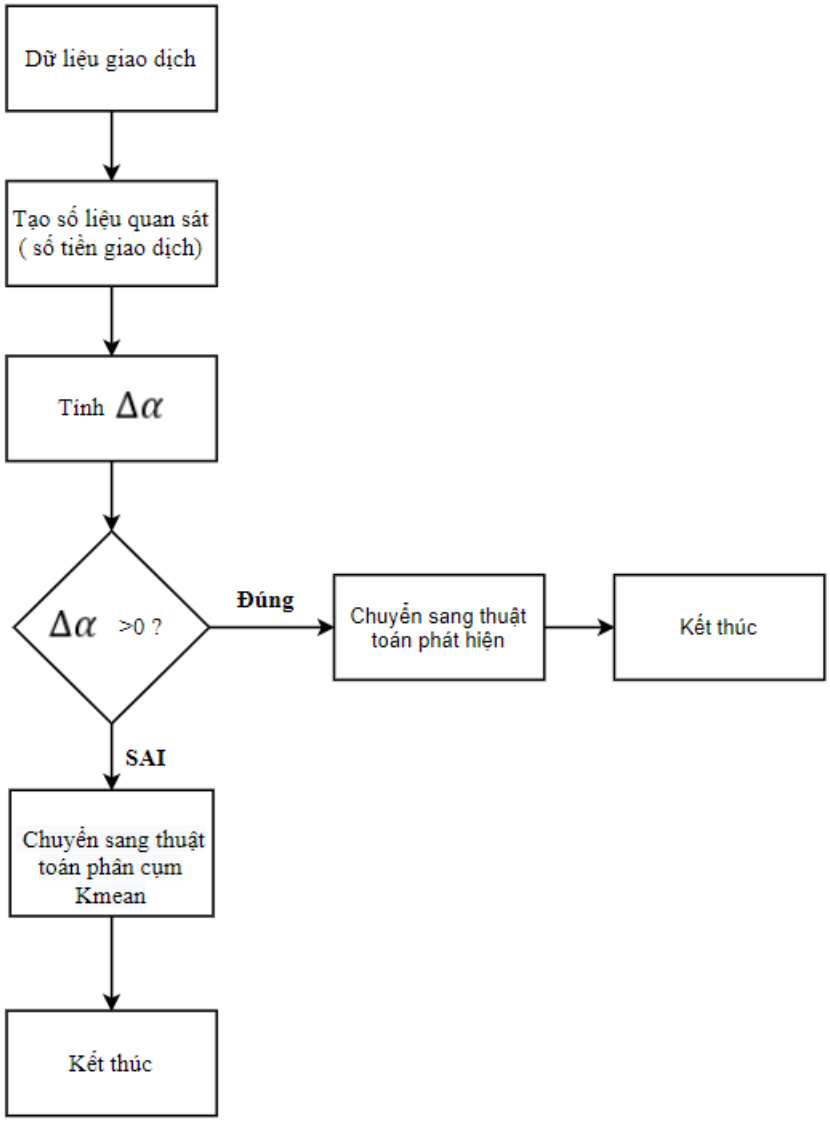
Bước 2: Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.

Bước 3: Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.

Bước 4: Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

4.2.2 Thuật toán đào tạo

Vấn đề tiếp theo là nếu giao dịch là bình thường, thì dữ liệu đó được đưa vào dùng để tính toán trong các lần giao dịch tiếp theo. Vì vậy, chúng ta cần có một bước gọi là đào tạo dữ liệu (training data). Với *thuật toán đào tạo* được thể hiện bằng sơ đồ bên dưới.



Hình 4.2: Thuật toán đào tạo

Các bước cụ thể như sau:

Bước 1: Từ dữ liệu giao dịch, xác định các tham số của mô hình và ký hiệu quan sát Q_{R+1}.

Bước 2: Tính α_1, α_2 và $\Delta\alpha$

Bước 3:

- Nếu $\Delta\alpha \leq 0$ thì giao dịch là bình thường, thêm dữ liệu và sử dụng thuật toán phân cụm K-mean để phân cụm với dữ liệu mới.
- Nếu $\Delta\alpha > 0$ thì giao dịch có thể là gian lân và chuyển sang thuật toán phát hiện.

4.2.3 Thuật toán phát hiện

Khi $\Delta\alpha > 0$ ta thực hiện *thuật toán phát hiện* như sau:

Bước 1: Nhập ngưỡng cho phép Z.

Bước 2 : Tính $\frac{\Delta\alpha}{\alpha_1}$

Bước 3:

- Nếu $\frac{\Delta\alpha}{\alpha_1} < Z$ thì sai khác là chấp nhận được và thêm dữ liệu vào để sử dụng thuật toán phân cụm K-mean.
- Nếu $\frac{\Delta\alpha}{\alpha_1} \geq Z$ thì hệ thống đưa ra cảnh báo giao dịch là gian lận.

Chú ý: Trong một số thử nghiệm với R trong khoảng từ 5 đến 25, với R =15 mang lại hiệu quả cao nhất [7]. Vì vậy, trong các phần tiếp theo, ta mặc định các số liệu có được khi áp dụng với R=15.

Chương trình python cho phương pháp có thể dowload tại:
https://github.com/haivodoimd12/HMM_ CFD.git

4.3 Nhận xét

4.3.1 Hiệu quả của phương pháp

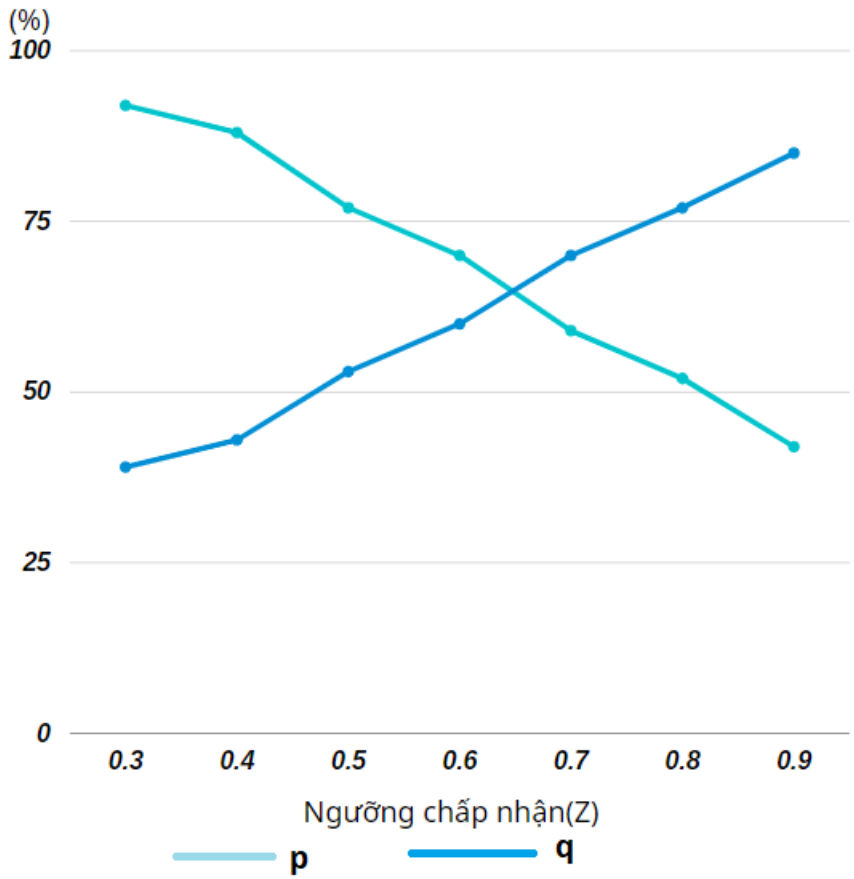
Đánh giá hiệu quả của phương pháp dựa trên việc chọn ngưỡng chấp nhận Z.
Ta thấy khi chọn Z càng nhỏ, số lượng các giao dịch bị cảnh báo gian lận càng nhiều. Tuy nhiên khi đó tỷ lệ chính xác sẽ rất thấp. Nếu chọn Z quá lớn sẽ dẫn tới việc một số giao dịch là gian lận nhưng không bị cảnh báo. Vì vậy chúng ta phải chọn ngưỡng chấp nhận Z sao cho hiệu quả. Ta tham khảo dữ liệu từ các thử nghiệm sau đây [9]

Trong đó:
p = (số gian lận phát hiện được)/(số gian lận thực tế)
q = Độ chính xác = (số gian lận)/(số cảnh báo gian lận)

Ngưỡng chấp nhận Z	$p(\%)$	$q(\%)$
0.3	92	39
0.4	88	43
0.5	77	53
0.6	70	60
0.7	59	70

Bảng 4.1: Hiệu quả phương pháp dựa trên ngưỡng chấp nhận

Dựa vào biểu đồ, ta thấy tùy thuộc vào mục tiêu và chiến lược trong thực tế, ta có thể chọn ngưỡng chấp nhận sao cho hợp lý. Ở đây chúng ta có thể chọn Z = 0,5 hoặc 0,6.



Hình 4.3: Hiệu quả phương pháp phụ thuộc ngưỡng chấp nhận Z

Độ chính xác thay đổi theo lượng dữ liệu giao dịch.
Cũng theo dữ liệu tham khảo từ những thử nghiệm [9] ta có bảng dữ liệu sau (với ngưỡng chấp nhận Z=0.5):

Số bộ dữ liệu	100	200	300	400	500	600
Độ chính xác	0.87	0.77	0.69	0.58	0.51	0.48

Bảng 4.2: Sự thay đổi của độ chính xác theo lượng dữ liệu giao dịch.

Ta thấy với lượng dữ liệu lớn, hiệu quả của phương pháp chưa cao.
Ta có một số nhận xét sau:

- Ưu điểm của phương pháp là đưa ra phán đoán nhanh vì các bước phân cụm dữ liệu, đào tạo dữ liệu đã được thực hiện trước khi giao dịch mới phát sinh.
- Nhược điểm của phương pháp: độ chính xác chưa cao, khả năng làm việc với dữ liệu lớn còn kém.

4.3.2 Cách khắc phục và phương hướng phát triển

- Thêm bước chọn k trong thuật toán phân cụm K-mean sao cho hiệu quả nhất.

- Thử nghiệm với dữ liệu thực tế để đưa ra ngưỡng chấp nhận hợp lý.

- Kết hợp và mở rộng với một số phương pháp khác như:

+ Sử dụng mô hình Markov ẩn phân kỳ[7].

+ Sử dụng mô hình bán Markov ẩn (SHMM) kết hợp với các nhân tố thông tin tiêu chuẩn (factorized information criterion – FIC) [9].

+ Sử dụng mô hình Markov ẩn đa phối cảnh (multi-perspective HMMs)[10].

Kết luận

Trong bài báo cáo, chúng ta đã trình bày được một số vấn đề sau:

1. Nhắc lại và trình bày một số kiến thức về xích Markov, quá trình Markov, mô hình Markov.
2. Trình bày định nghĩa và một số tính chất đặc trưng của mô hình Markov ẩn cùng ví dụ minh họa.
3. Trình bày ba bài toán cơ bản của mô hình Markov ẩn và một số ứng dụng trong thực tế.
4. Trình bày phương pháp phát hiện gian lận thẻ tín dụng bằng cách sử dụng mô hình Markov ẩn bao gồm: các thuật toán, nhận xét hiệu quả phương pháp cùng phương hướng khắc phục, phát triển.

Tài liệu tham khảo

[1] T. N. report (2017). Card & Mobile payment Industry Statistics. [Online]. Available: https://nilsonreport.com/publication _ chart_ and_ graphs_ archive.php?1=1& year=2017.

[2] The Nilson Report (2017). [Online]. Available: https://nilsonreport.com/upload/content_ promo/The_ Nilson_ Report_ Issue_ 1118.pdf

[3] S. al, (2015). Credit Card Fraud Detection Using Decision Tree Induction Algorithm. International Journal of Computer Science and Mobile Computing, vol. 4, no. 4.

[4] Raghavendra Patidar, Lokesh Sharma (2011). Credit Card Fraud Detection Using Neural Network. International Journal of Soft Computing and Engineering (IJSCE), vol. 1.

[5] Dinesh L. Talekar, K. P. Adhiya (2014). Credit Card Fraud Detection Using Decision Tree For Tracing Email And Ip. IJCSI International Journal of Research(IJMER), vol. 4, no. 9

[6] Mr.P.Matheswaran, Mrs.E.SivaSankari ME, Mr.R.Rajesh (2015). Fraud Detection in Credit Card Using Data Mining Techniques. IJRSET, vol. II, no. I.

[7] William N. Robinson & Andrea Aria, Sequential fraud detection for prepaid cards using hidden Markov model divergence, Computer Information Systems, Georgia State Unvisersity, Atlanta, GA, USA [Online] https://www.sciencedirect.com/science/article/abs/pii/S0957417417305894

[8] Nguyễn Văn Hiếu, Thuật toán phân cụm k-means và code minh họa bài toán phân cụm, https://nguyenvanhieu.vn/thuat-toan-phan-cum-k-means

[9] A.Prakash & C.Chandrasekar, A Novel Hidden Markov Model for Credit Card Fraud Detection, International Journal of Computer Applications (0975 – 8887) Volume 59– No.3, 2012.

[10] Lucas & Pierre-Edouard Portier a, Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs, [Online] https://www.sciencedirect.com/science/article/pii/S0167739X19300664