MỘT SỐ THÍ NGHIỆM VỀ XÁC SUẤT THỐNG KÊ ỨNG DỤNG TRONG LĨNH VỰC BÓNG ĐÁ

Thái Bình Dương
Khoa Hệ thống thông tin
Trường Đại học Công nghệ thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh
23520356@gm.uit.edu.vn

***Tóm tắt*— Đồ án này nghiên cứu ứng dụng các lý thuyết xác suất thống kê vào lĩnh vực bóng đá thông qua ba bài toán cụ thể. Đầu tiên, mô phỏng các tình huống đá phạt đền sử dụng sinh số ngẫu nhiên để phân tích xác suất thành công của cầu thủ. Tiếp theo, kiểm định mối quan hệ giữa kết quả trận đấu của đội ghi bàn mở tỷ số và chất lượng giải đấu bằng các phương pháp kiểm định Chi bình phương. Cuối cùng, dự đoán kết quả của một trận đấu dựa trên dữ liệu lịch sử bằng xích Markov, đánh giá bằng ba giả thiết khác nhau. Kết quả nghiên cứu không chỉ minh họa vai trò quan trọng của xác suất thống kê mà còn cho thấy các ứng dụng thực tiễn trong dự đoán và phân tích dữ liệu bóng đá.**

Từ khoá—Bóng đá, Toán ứng dụng, Ma trận, Đại số tuyến tính, Xác suất thống kê, Phương pháp Monte-Carlo, Sinh số ngẫu nhiên, Kiểm định độc lập, Xích Markov, Quá trình ngẫu nhiên

I. GIỚI THIỆU ĐỒ ÁN

Bóng đá được ví như môn thể thao "vua" do đã có lịch sử lâu đời và tính chất hấp dẫn trong các trận đấu. Sự phát triển và phổ biến của bóng đá cũng đặt ra nhiều câu hỏi về các số liệu ở lĩnh vực này. Trong những năm qua, có khá nhiều nghiên cứu toán học và học máy về bóng đá đã được thực hiện như: dự đoán kết quả trận đấu [1]; tính xác suất của thủ môn bắt được bóng trong lượt đá phạt đền [2, 3]; hay dự đoán số bàn thắng mong đợi của một đội bóng trong trận đấu [4].

Với mong muốn vận dụng kiến thức đã học về xác suất thống kê chuyên sâu và niềm đam mê với bóng đá, em đã thực hiện đồ án này. Đồ án tập trung vào ba bài toán: Mô phỏng đá phạt đền bằng sinh số ngẫu nhiên (mục II); Kiểm định tính độc lập giữa kết quả một trận đấu của đội ghi bàn mở tỷ số với giải đấu (mục III); Dự đoán kết quả trận đấu thứ N của một đội bóng bằng xích Markov (mục IV). Những bài toán trên không chỉ giúp giải quyết các vấn đề lý thú trong bóng đá mà còn minh họa ứng dụng thực tiễn của xác suất thống kê và các mô hình toán học.

II. MÔ PHỎNG ĐÁ PHẠT ĐỀN BẰNG SINH SỐ NGẪU NHIÊN

A. Giới thiệu bài toán

Trong phần này, em sẽ thực hiện thí nghiệm mô phỏng việc sút phạt đền sử dụng phương pháp sinh số ngẫu nhiên. Cho rằng hướng sút của cầu thủ và hướng đổ người của thủ môn là hai yếu tố độc lập, với xác suất sút vào khung thành của tất cả cầu thủ và xác suất bắt được bóng của tất cả thủ môn đều như nhau và biến cố là rời rạc [5]. Dựa trên quy ước lịch sử các cú sút và hướng đổ người thành các vùng cố định, em sẽ thực hiện sinh số ngẫu nhiên và mô phỏng việc một cầu thủ thực hiện 10, 50, 100 và 1000 cú sút phạt đền. Sau đó, thực hiện đánh giá tổng quan phương pháp với số liệu đã được kiểm chứng để tìm những hạn chế để cải thiện.

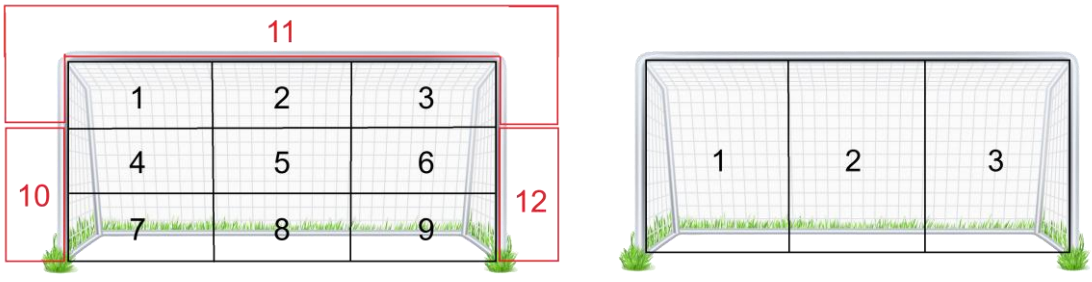
B. Dữ liệu

Để thực hiện bài toán mô phỏng, em sử dụng bộ dữ liệu về sút luân lưu trong các trận đấu thuộc khuôn khổ giải bóng đá World Cup [6]. Bộ dữ liệu này gồm thông tin các hướng sút và hướng đổ người của thủ môn của 320 lượt đá phạt đền kể từ năm 1982. Dữ liệu đã được em cập nhật đến năm tổ chức gần nhất (2022).

Bộ dữ liệu gốc quy ước hướng sút thành 9 vùng nhỏ chia đều trong khung thành (thứ tự từ trái sang phải và từ trên xuống dưới) và hướng đổ người được chia thành 3 vùng dọc theo khung thành. Để xét luôn đến các trường hợp cầu thủ có sút theo hướng đó nhưng bóng ra ngoài (OnTarget có giá trị là 0), em đã thực hiện xử lý dữ liệu với cột "Zone". Cụ thể, các trường hợp sút ra ngoài theo hướng của vùng 4 và 7 thì Zone bằng 10, sút ra ngoài theo hướng của vùng 1, 2 và 3 thì Zone bằng 11, và sút ra ngoài theo hướng của vùng 6 và 9 thì Zone bằng 12. Quy ước hướng sút của cầu thủ và hướng đổ người của thủ môn được minh hoạ cụ thể trong *Hình 1*.

C. Phương pháp Monte-Carlo và sinh số ngẫu nhiên

Phương pháp Monte-Carlo là phương pháp mô phỏng bằng xác suất để dự đoán các kết quả có thể xảy ra của một sự kiện không chắc chắn do có sự can thiệp của các yếu tố ngẫu nhiên. Việc mô phỏng được thực hiện bằng cách tạo các dữ liệu là các số giả ngẫu nhiên. Các số giả ngẫu nhiên này được sinh ra từ một hàm mật độ xác suất (PDF) hoặc hàm khối lượng xác suất (PMF) xây dựng bởi các dữ liệu lịch sử hoặc đánh giá chủ quan phân phối xác suất của các biến độc lập. Thực hiện mô phỏng này nhiều lần, liên tục tạo ra các giá trị ngẫu nhiên của biến độc lập, cho đến khi thu thập đủ kết quả để tạo thành một mẫu đại diện cho số lượng gần như vô hạn các kết hợp có thể có [7].



(a) (b)

Hình 1: Quy ước hướng sút của cầu thủ sút phạt đền (a) và hướng đổ người của thủ môn (b)

Bảng I: Bảng phân phối xác suất về hướng đổ người của thủ môn

X	**1**	**2**	**3**
P	0.471875	0.115625	0.40625

Bảng II: Bảng phân phối xác suất về hướng sút của cầu thủ

X	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
P	0.071875	0.04375	0.04375	0.125	0.065625	0.1125	0.221875	0.075	0.153125	0.03125	0.046875	0.009375

Ở bài toán này, do cú sút và hướng đổ người được xem là các biến cố rời rạc (mục II.A), em sẽ xây dựng hàm PMF của hướng sút bóng và hướng đổ người.

D. Thực hiện thí nghiệm

Sử dụng ngôn ngữ lập trình Python cùng các công cụ và thư viện cần thiết, em đã thực hiện xây dựng PMF của hướng sút và hướng đổ người dựa trên bộ dữ liệu ban đầu. Phân phối xác suất của cú sút và hướng đổ người của dữ liệu này được ghi rõ ở Bảng II và Bảng I, với **X** là giá trị và **P** là xác suất. Sau đó, tiến hành biến đổi bảng phân phối PMF về hàm phân phối tích luỹ, và xây dựng hàm sinh số ngẫu nhiên dựa trên việc biến đổi hàm ngược.

Biểu thức tổng quát cho việc biến đổi hàm ngược của hàm phân phối tích luỹ $F_X(x)$ là:

$$g(u) = F_X^{-1}(u) \quad (1)$$

Trong đó giá trị u là giá trị được sinh ngẫu nhiên tuân theo phân phối đều: $u \sim Uniform(0,1)$.

Từ (1) em sẽ suy ra được hàm sinh số ngẫu nhiên $g_1(u)$ về hướng sút:

$$g_1(u) = \begin{cases} 1 & 0 \le u \le 0.071875 \\ 2 & 0.071875 < u \le 0.115625 \\ 3 & 0.115625 < u < 0.159375 \\ 4 & 0.159375 < u \le 0.284375 \\ 5 & 0.284375 < u \le 0.35 \\ 6 & 0.35 < u \le 0.4625 \\ 7 & 0.4625 < u \le 0.684375 \\ 8 & 0.684375 < u \le 0.753975 \\ 9 & 0.753975 < u \le 0.9125 \\ 10 & 0.9125 < u \le 0.94375 \\ 11 & 0.94375 < u \le 0.990625 \\ 12 & 0.990625 < u \le 1 \end{cases}$$

và hàm $g_2(u)$ với hướng đổ người:

$$g_2(u) = \begin{cases} 1 & u \le 0.471875 \\ 2 & 0.471875 < u \le 0.5875 \\ 3 & 0.5875 < u \le 1 \end{cases}$$

Cuối cùng, xây dựng mã nguồn sinh số ngẫu nhiên để ra kết quả. Quy tắc xét kết quả đá phạt đền như sau:

- Nếu hướng sút sinh ra số 10, 11, 12, thì kết quả là "Sút ra ngoài" với mọi trường hợp thủ môn đổ người.
- Nếu hướng sút trùng với quy ước vùng bắt bóng của thủ môn thì kết quả là "Bị cản phá". Ví dụ, nếu hướng sút là 1, 4, hoặc 7 mà thủ môn đổ người ở hướng 1 thì sẽ tính là "Bị cản phá". Các trường hợp thủ môn đổ người ở hướng 2 và 3 sẽ xét tương tự theo quy chiếu ở *Hình 1*.
- Các kết quả còn lại được tính là "Ghi bàn".

Thực hiện giả lập một cầu thủ sút 10, 50, 100 và 1000 quả phạt đền liên tục, kết quả ghi nhận được trình bày ở Bảng III. Các giá trị ở Bảng III cho ta thấy được rằng mặc dù khi thực hiện càng nhiều lần sút, tỉ lệ chính xác không có sự thay đổi quá lớn. Số cú sút ghi bàn chỉ chiếm khoảng 50-60% tổng số cú sút ở mỗi lần thực hiện mô phỏng.

E. Đánh giá, nhận định kết quả

Tuy nhiên, nếu xét trên số liệu đã được tính toán và kiểm chứng thì tỉ lệ thực hiện cú sút thành công sẽ rơi vào khoảng 70-80% [3], nghĩa là tỉ lệ được mô phỏng vẫn còn nhiều bất cập. Một số yếu tố đã được em xem xét đến, trong đó có hai nguyên nhân được em quan tâm nhất.

Thứ nhất, em cho rằng là thiếu sót ở bộ dữ liệu gốc, đặc biệt là dữ liệu hướng đổ người của thủ môn. Nếu như dữ liệu sút của cầu thủ được chia thành 9 phần, sau được em mở rộng lên 12 phần, thì dữ liệu của thủ môn chỉ được chia thành 3 phần. Điều này đồng nghĩa với việc khi sinh số ngẫu nhiên với các trường hợp ghi bàn và bị cản phá, thì thực tế chỉ có tối đa là 9 trường hợp được xét, vì quy ước 1 vùng đổ người của thủ môn sẽ tương đương 3 vùng sút của cầu thủ (việc chia trường hợp đã được đề cập ở phần D của bài toán này). Chưa kể, xác suất ghi bàn với mỗi vùng sút lại không giống nhau, nên khó phản ánh khách quan.

Thứ hai, có thể một số yếu tố độc lập khác chưa được xét đến, ví dụ như trường hợp cầu thủ không sút ra ngoài và thủ môn đổ đúng hướng nhưng bắt không dính bóng hoặc không cản phá được do bóng nhanh hoặc vị trí bóng hiểm (mặc dù trường hợp này khá ít và có thể bỏ qua). Ngoài ra, trong các loạt sút luân lưu, thứ tự sút cũng có thể ảnh hưởng đến xác suất ghi bàn của cầu thủ [8].

Nhìn chung, việc mô phỏng đá phạt đền bằng sinh số ngẫu nhiên là hoàn toàn khả thi dù có hạn chế khách quan. Những hạn chế này đã giúp cung cấp cho em cái nhìn sâu sắc hơn về các yếu tố tiềm tàng có thể ảnh hưởng đến kết quả ở bài toán này, đồng thời chỉ ra các điểm cần cải thiện trong phương pháp thực hiện. Trong tương lai, em sẽ xây dựng các phương pháp mở rộng để xử lý các trường hợp mất cân bằng dữ liệu giữa thủ môn và cầu thủ cũng như các yếu tố khác như đã đề cập ở trên.

III. KIỂM ĐỊNH TÍNH ĐỘC LẬP GIỮA KẾT QUẢ TRẬN ĐẤU CỦA ĐỘI BÓNG MỞ TỶ SỐ VỚI GIẢI ĐẤU

A. Giới thiệu bài toán

Trong phần này, em sẽ thực hiện kiểm định tính độc lập giữa kết quả trận đấu của đội bóng mở tỷ số với chất lượng của giải đấu. Số liệu đã được khảo sát tại bốn giải đấu bóng đá vô địch quốc gia, với hơn 1000 trận đấu được khảo sát. Phương pháp được sử dụng trong bài toán này là thực hiện kiểm định Chi bình phương.

Bảng III: Kết quả thí nghiệm bài toán đá phạt đền

Ghi bàn	5	27	58	573
Bị cản phá	3	20	34	338
Sút ra ngoài	2	3	8	122
Tổng	**10**	**50**	**100**	**1000**

B. Dữ liệu

Dữ liệu được sử dụng là kết quả toàn bộ các trận đấu có bàn thắng của bốn giải vô địch quốc gia hàng đầu châu Âu bao gồm: Premier League (Anh), Bundesliga (Đức), La Liga (Tây Ban Nha) và Serie A (Ý). Ba yếu tố được thống kê là số trận đấu mà đội bóng mở tỷ số giành chiến thắng, số trận đấu đội bóng mở tỷ số nhưng thua cuộc (thắng lội ngược dòng), và số trận hoà có bàn thắng. Toàn bộ số liệu được thu thập thủ công và kết quả được trình bày ở Bảng IV.

C. Kiểm định độc lập bằng kiểm định Chi bình phương

Kiểm định Chi bình phương (Chi-Square Test) là một phương pháp thống kê được sử dụng để kiểm tra sự khác biệt giữa dữ liệu thực nghiệm và dữ liệu lý thuyết dự kiến. Phương pháp này thường sử dụng để kiểm tra mối quan hệ (độc lập hoặc phụ thuộc) giữa hai đại lượng ngẫu nhiên X và Y bất kỳ. Chúng ta đặt giả thiết cần kiểm định (giả thiết không):

$$H_0: \ X \text{ và } Y \text{ độc lập với nhau}$$

và đối thiết của nó:

$$H_1: \ X \text{ và } Y \text{ phụ thuộc lẫn nhau}$$

ở mức ý nghĩa α.

Gọi $x_1, x_2, \ldots, x_k$ là mẫu ngẫu nhiên từ đại lượng X, và $y_1, y_2, \ldots, y_l$ là số liệu quan sát đại lượng Y trên đại lượng X. Chúng ta có thể quy bài toán so sánh l bộ số liệu Y trên k mẫu X ở trên thành bài toán so sánh bộ số liệu gồm $k \times l$ kích thước mẫu (số liệu hai chiều) với một bộ số liệu lý thuyết. Muốn vậy, ta phải tính tổng của mỗi cột x_i (C_i), tổng của mỗi hàng y_j (H_j), và tổng số mẫu được khảo sát (**N**) ở Bảng V.

Sau đó, thực hiện tính toán kiểm định bằng công thức (2):

$$Q = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}} \tag{2}$$

với n_{ij} là tần số quan sát, n'_{ij} là tần số kì vọng:

$$n'_{ij} = \frac{\mathrm{H}_i \mathrm{C}_j}{\mathrm{N}} \tag{3}$$

Bảng V: Minh hoạ kích thước mẫu hai chiều

X / **Y**	x_1	x_2	…	x_k	**H**
y_1	n_{11}	n_{12}	…	n_{1k}	H_1
y_2	n_{21}	n_{22}	…	n_{2k}	H_2
…	…	…	…	…	…
y_l	n_{l1}	n_{l2}	…	n_{lk}	H_l
C	C_1	C_2	…	C_k	**N**

Bảng IV: Thống kê kết quả của đội bóng mở tỷ số ở các giải đấu

Giải đấu / **Kết quả**	EPL	Laliga	Serie A	Budesliga
Mở tỷ số và thắng	236	241	222	189
Hoà	70	75	82	68
Mở tỷ số nhưng thua	59	38	36	37

Giả sử nếu H_0 đúng thì phân phối của Q sẽ gần đúng với phân phối Chi bình phương:

$$Q \sim \chi^2_{1-\alpha}(df)$$

với độ tin cậy $1-\alpha$ [9] và bậc tự do như sau:

$$df = (k-1)(l-1) \tag{4}$$

trong đó k là số cột, l là số hàng.

Như vậy, chúng ta sẽ phải so sánh giá trị của Q với giá trị tới hạn $\chi^2_{1-\alpha}(df)$ trong bảng phân phối Chi bình phương:

- Nếu $Q < \chi^2_{1-\alpha}(df)$: Chấp nhận H_0.
- Nếu $Q \geq \chi^2_{1-\alpha}(df)$: Bác bỏ H_0, chấp nhận H_1.

Giá trị bậc tự do (df) cũng tương đương với sự khác biệt giữa các tham số của giả thiết và từng đối thiết [10].

D. Thực hiện thí nghiệm

Việc thu thập dữ liệu được tiến hành thủ công bằng cách xem kết quả chi tiết của 1353 trận đấu thuộc bốn giải đấu đã được đề cập trong mục B của bài toán này. Sau khi hoàn thành việc thu thập dữ liệu, em đã sử dụng ngôn ngữ lập trình Python với hai thư viện chính là `numpy` và `scipy.stats` để xử lý và tính toán kết quả.

Đầu tiên, em chuyển đổi dữ liệu quan sát của số liệu có trong Bảng IV thành tần số kỳ vọng (Bảng VI). Kết quả kì vọng được làm tròn đến chữ số thập phân thứ hai.

Gọi:

- H_0: Kết quả của đội bóng ghi bàn mở tỷ số độc lập với giải đấu.
- H_1: Kết quả của đội bóng ghi bàn mở tỷ số phụ thuộc vào giải đấu.

Sau đó, thực hiện tính giá trị của Q và giá trị tới hạn $\chi^2_{1-\alpha}(df)$ như theo lý thuyết đã được đề cập trong mục C. Với $k = 4,\ l = 3,$ bậc tự do của số liệu trong Bảng IV theo công thức (4) sẽ là:

$$df = (k-1)(l-1) = (4-1)(3-1) = 6$$

Về mức ý nghĩa α của bài toán, em lựa chọn giá trị là 5% vì đây là mức truyền thống [11]. Chưa kể, phần lớn nghiên cứu thường chọn mức ý nghĩa cho sai lầm loại I (sai lầm tích cực) là khoảng 1-10% [12], nên việc chọn 5% là mức bình quân, vừa đủ nhưng phù hợp nhất.

Cuối cùng, chạy mã nguồn để tính toán, kết quả cuối cùng được đưa ra là:

$$Q = 8.515838;\ \chi^2_{0.95}(6) = 12.59159$$

Vì $Q < \chi^2_{0.95}(6)$ nên chấp nhận giả thuyết H_0. Vậy kết quả của đội bóng mở tỷ số độc lập với giải đấu.

Bảng VI: Kết quả kì vọng được chuyển đổi từ Bảng IV

Giải đấu / Kết quả	EPL	Laliga	Serie A	Budesliga
Mở tỷ số và chiến thắng	239.56	232.34	223.15	192.96
Hoà	79.52	77.18	74.13	64.10
Mở tỷ số nhưng thua	45.86	44.45	42.72	36.94

E. Đánh giá, nhận định kết quả

Việc thực hiện bài toán đã đảm bảo đúng và đầy đủ các yêu cầu về kiến thức trong môn học và mục tiêu đề ra của người thực hiện đồ án. Thí nghiệm đã theo sát với cơ sở lý thuyết về kiểm định Chi bình phương đã được nêu chi tiết tại mục C của bài toán này. Kết quả cuối cùng được tính toán cũng đã kiểm định chính xác với các giả thuyết mong đợi mà người thực hiện đồ án đã đặt ra, rằng yếu tố kết quả trận đấu của đội ghi bàn mở tỷ số trước và yếu tố về chất lượng giải đấu là hoàn toàn độc lập.

Phương pháp kiểm định Chi bình phương cũng là phương pháp tin cậy để thực hiện các bài toán kiểm định tính độc lập với trường hợp các biến ngẫu nhiên có tần số kỳ vọng cao.

IV. Dự đoán kết quả trận đấu của một đội bóng bằng Xích Markov

A. Giới thiệu

Trong phần này, em sẽ thực hiện bài toán về việc dự đoán kết quả trận đấu thứ *N* của một đội bóng bằng xích Markov. Trên thực tế, bài toán dự đoán kết quả bằng xích Markov cũng được áp dụng trong nhiều môn thể thao khác như bóng chày [13], cricket [14] hay khúc côn cầu [15].

Sử dụng dữ liệu lịch sử thi đấu được ghi nhận trước đó, em sẽ tiến hành xây dựng ma trận chuyển đổi trạng thái từ kết quả của trận trước sang kết quả của trận sau. Bài toán sẽ thử nghiệm với kết quả thi đấu của một câu lạc bộ bóng đá bất kì, thực hiện thí nghiệm và đánh giá ba giả thiết: dự đoán nếu biết được kết quả trận trước đó (thắng, hoà, thua), dự đoán nếu cho phân bố ban đầu là xác suất kết quả trong lịch sử (mùa giải trước đó), và dự đoán nếu cho rằng khả năng giành chiến thắng, hoà hoặc thua cuộc là bằng nhau.

B. Dữ liệu

Để thực hiện bài toán này, em đã sử dụng bộ dữ liệu về kết quả các trận đấu trong khuôn khổ giải bóng đá Ngoại hạng Anh ở mùa giải 2023/2024 [16]. Bộ dữ liệu gồm toàn bộ thông tin của 380 trận đấu, với mỗi trận đấu được ghi nhận hai lần (thông tin được ghi nhận thành hai kiểu là thi đấu trên sân nhà và sân khách, thay đổi ở cột "Venue"). Tổng cộng có tất cả 760 dòng dữ liệu. Trong bài toán này, em sẽ chỉ sử dụng dữ liệu của câu lạc bộ bóng đá Machester United để thử nghiệm, tức là sẽ chỉ có 38 dòng dữ liệu được sử dụng.

C. Xích Markov và ma trận chuyên đổi trạng thái

Xét một quá trình ngẫu nhiên $\{X_0, X_1, X_2, ..., X_n\}$ với thời gian rời rạc có không gian trạng thái Z là một tập hữu hạn hoặc vô hạn đếm được. Quá trình này sẽ được gọi là xích Markov nếu nó thoả mãn tính Markov với mọi vectơ $x_0, x_1, x_2, ...x_{n+1}$, trong đó $x_k \in Z \ (k \le n)$, $n \ge 0$ theo công thức (5):

$$\begin{aligned} &P\left(X_{n+1} = x_{n+1} \mid X_0 = x_0, X_1 = x_1, ...X_n = x_n\right) \\ &= P\left(X_{n+1} = x_{n+1} \mid X_n = x_n\right) \end{aligned} \quad (5)$$

Nếu ta coi thời điểm *n* là hiện tại và thời điểm *n* + 1 là tương lai, thì ta có thể coi *n* – 1 thời điểm trước đó là quá khứ. Nói một cách dễ hiểu cho công thức (5), thì phân phối xác suất có điều kiện trong trạng thái tương lai sẽ chỉ phụ thuộc vào hiện tại chứ không phụ thuộc với quá khứ của nó [9]. Tại thời điểm trạng thái *i* có tương lai là *j*, dựa theo công thức (5) chúng ta có thể viết thành:

$$p_{ij} = P\left(X_{n+1} = j \mid X_n = i\right) \quad (6)$$

với p_{ij} được gọi là xác suất chuyển một bước, hoặc xác suất chuyển từ trạng thái *i* sang trạng thái *j*. Các xác suất chuyển một bước sẽ có tính chất sau:

$$0 \le p_{ij} \le 1; \ \sum_j p_{ij} = 1; \ i, j = 0, 1, 2, \ldots$$

Nếu p_{ij} phụ thuộc vào *n* thì xích Markov sẽ không thuần nhất, ngược lại nếu không phụ thuộc thì xích Markov sẽ thuần nhất theo thời gian. [17]. Các xác suất chuyển một bước của xích Markov sẽ tạo nên một ma trận được gọi là ma trận xác suất chuyển (hoặc ma trận chuyển, ma trận chuyển đổi trạng thái), ký hiệu là *P* [9]:

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \ldots \\ p_{10} & p_{11} & p_{12} & \ldots \\ \vdots & \vdots & \vdots & \ldots \\ p_{i0} & p_{i1} & p_{i2} & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

Nếu thực hiện *m* bước chuyển trạng thái *j* từ trạng thái *i*, ta có thể viết là $P^{(m)}$, hay viết dưới dạng tích ma trận là P^m. Nói cách khác, ma trận chuyển xác suất sau *m* bước là tích *m* lần của xác suất chuyển một bước *P*.

Cũng ở công thức (6), chúng ta cũng cần phải để ý đến trạng thái tại thời điểm ban đầu của nó, tức là trạng thái *i* và chưa chuyển sang bất kì trạng thái khác. Khi đó, chúng ta gọi đó là phân phối ban đầu (hoặc vectơ trạng thái ban đầu), kí hiệu là π_0:

$$\pi_0 = p_0 = P(X_0 = i); \sum p_0 = 1$$

Như vậy, phân phối của xích Markov tại bước thứ *m* (hoặc vectơ trạng thái thứ *m*) kí hiệu sẽ là π_m, sẽ là tích của phân phối ban đầu và ma trận chuyển sau *m* bước:

$$\pi_m = \pi_0 P^m \quad (7)$$

và từ (7) ta có thể suy ra được phân phối của xích Markov tại trạng thái thứ *m* + 1 sẽ là phân phối tại bước *m* nhân với ma trận chuyển đổi trạng thái.

$$\pi_{m+1} = \pi_0 P^{m+1} = \pi_0 P^m P = \pi_m P \quad (8)$$

Nếu như phân phối tại (8) không thay đổi trạng thái sau một lượng lớn thời kỳ (hay lượng lớn bước), thì phân phối sẽ đạt trạng thái cân bằng, được gọi là phân phối dừng (hoặc phân bố dừng) [18]. Lúc này phân phối đạt trạng thái ổn định.

Một số tài liệu hoặc nghiên cứu sẽ xét tổng xác suất trạng thái các cột bằng 1, thay vì tổng xác suất trạng thái các hàng bằng 1 [19]. Nếu quy ước tổng xác suất trạng thái các cột bằng 1 sẽ cần phải thực hiện chuyển vị phân phối ban đầu và ma trận trạng thái, hoặc sẽ thực hiện nhân giao hoán ma trận

Bảng VII: Tỉ lệ kết quả trận đấu so với trận trước đó

Trận sau / Trận trước	Thắng	Hoà	Thua
Thắng	0.411765	0.117647	0.470588
Hoà	0.333333	0.166667	0.500000
Thua	0.571429	0.214286	0.214285

chuyển với phân phối ban đầu. Sau cùng, cách làm này đưa về dạng tương tự như xét tổng các hàng bằng 1, nên việc dùng công thức (7) và (8) được coi là phù hợp và ngắn gọn.

D. Thực hiện thí nghiệm

Sử dụng ngôn ngữ lập trình Python với thư viện `numpy` và `pandas`, em đã thực hiện thu thập, xử lý dữ liệu và thực hiện xây dựng ma trận chuyển đổi với ba trạng thái: thắng, hoà, thua. Thực hiện chia tỉ lệ, kết quả được ghi ở Bảng VII. Gắn nhãn 0 là các trận thắng (W), 1 là các trận hoà (D) và 2 là các trận thua (L). Từ số liệu ở Bảng VII chúng ta có ma trận chuyển đổi trạng thái như sau:

$$P = \begin{pmatrix} 0.411765 & 0.117647 & 0.470588 \\ 0.333333 & 0.166667 & 0.500000 \\ 0.571429 & 0.214286 & 0.214285 \end{pmatrix}$$

Ma trận trên sẽ tương ứng với đồ thị trạng thái ở *Hình 2*. Như đã đề cập ở phần A của bài toán này, em sẽ thực hiện thí nghiệm với ba trường hợp.

1) Giả thiết 1: Phân bố ban đầu dựa trên kết quả của trận đấu trước đó

Giả sử như kết quả trận trước đó là thắng, phân bố xác suất ban đầu sẽ là:

$$\pi_0(\mathrm{W}) = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$

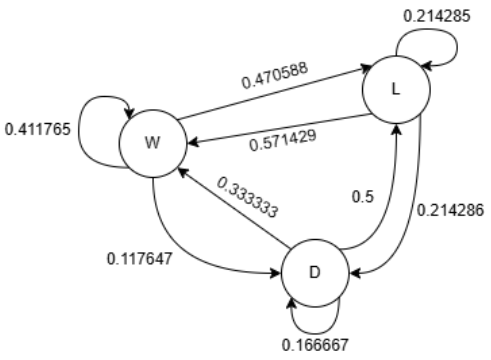
và tương tự cho phân bố ban đầu nếu như trận trước đó là thua và hoà:

$$\pi_0(\mathrm{D}) = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}; \pi_0(\mathrm{L}) = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$$

Chạy mã nguồn giả lập phân bố xác suất sau trận đầu tiên cho cả ba trường hợp (kết quả ghi nhận ở Bảng VIII), ta thấy rằng nếu trận trước của đội bóng là hoà và thắng, khả năng cao trận tiếp theo sẽ là trận thắng (với xác suất gần 48%). Tuy nhiên, nếu trận trước đó là trận thua, xác suất trận sau đó là trận thắng hoặc trận thua là gần tương đương nhau.

Tiếp tục giả lập thêm nhiều lần, ta thấy cả ba trường hợp sẽ đều đạt phân bố dừng ở trạng thái thứ 14 (trận thứ 14):

$$\pi = \begin{pmatrix} 0.45946 & 0.16216 & 0.37838 \end{pmatrix}$$



Hình 2: Đồ thị chuyển trạng thái kết quả trận đấu

Bảng VIII: Phân bố kết quả trận đấu sau giả lập trận thứ nhất

Xác suất / Phân phối đầu	Thắng	Hoà	Thua
Thắng	0.477673	0.168891	0.353435
Hoà	0.478525	0.174136	0.347339
Thua	0.429172	0.148860	0.421969

Điều này đồng nghĩa với việc mô hình đã đạt trạng thái ổn định sau 14 trận đấu.

2) Giả thiết 2: Phân bố ban đầu là phân bố kết quả của mùa giải trước

Thống kê lịch sử thi đấu 38 trận ở mùa giải trước của đội bóng, có tất cả 18 trận thắng, 6 trận hoà và 14 trận thua. Như vậy, phân bố ban đầu của đội bóng dựa trên lịch sử thi đấu H sẽ là:

$$\pi_0(\mathrm{H}) = \begin{pmatrix} 0.473684 & 0.157895 & 0.368421 \end{pmatrix}$$

Chạy giả lập với trường hợp này, sau trận đấu thứ nhất thì kết quả có sự biến động, khi xác suất thắng có xu hướng giảm, còn xác suất hoà và thua có xu hướng tăng, nhưng sự biến động này được em đánh giá là không quá nhiều. Trường hợp này cũng sẽ đạt phân bố dừng tương đương ở giả thiết 1):

$$\pi = \begin{pmatrix} 0.45946 & 0.16216 & 0.37838 \end{pmatrix}$$

tuy nhiên, phân bố đạt trạng thái dừng tại bước thứ 11 (trận đấu thứ 11), nghĩa là mô hình sẽ có tính ổn định tốt hơn nếu dùng giả thuyết này.

3) Giả thiết 3: Tỉ lệ thắng, hoà, thua là như nhau

Giả sử như xác suất thắng, hoà, thua của đội bóng là bằng nhau, tức là đội bóng chưa đá hoặc không có lịch sử thi đấu rõ ràng, khi này chúng ta sẽ có phân phối ban đầu là:

$$\pi_0(\mathrm{U}) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

với U ở đây là trường hợp chưa xác định trạng thái trước đó (Undefined).

Chạy giả lập với trường hợp này, phân phối xác suất có sự biến động mạnh sau bước đầu tiên. Xác suất hoà trận đấu giảm gần một nửa (còn lại 16.62%), trong khi khả năng thua và thắng tăng so với ban đầu, lần lượt là khoảng 20% và 30%. Thực hiện chạy nhiều lần, phân bố đạt ổn định tại bước thứ 13 (trận đấu thứ 13), với phân bố cũng giống như hai giả thiết trước đó:

$$\pi = \begin{pmatrix} 0.45946 & 0.16216 & 0.37838 \end{pmatrix}$$

E. Đánh giá, nhận định kết quả

Mặc dù thực hiện xích Markov với phân bố dừng là ba giả thiết khác nhau, nhưng kết quả của cả ba là gần giống nhau, khi đều cùng hội tụ về một phân bố dừng. Tuy nhiên, số bước để chuyển về trạng thái ổn định lại có sự khác biệt. Và qua ba giả thiết thử nghiệm ở mục D của bài toán này, chúng ta cũng để ý rằng, nếu phân phối đầu dựa trên lịch sử của càng nhiều trận đấu, hay biết rõ hiện tại dựa trên quá khứ trước đó (như ở giả thiết 2 là dựa trên lịch sử của cả mùa giải – tức khoảng 38 trận) thì kết quả dự đoán cho tương lai sẽ ổn định tốt hơn. Điều khá thú vị là việc giả sử không có thông tin lịch sử thi đấu rõ ràng thì kết quả lại ổn định hơn so với việc đã biết lịch sử thi đấu của trận trước đó.

Việc thực hiện bài toán đã bám sát cơ sở lý thuyết của xích Markov. Mô hình Markov cũng là mô hình tối ưu để có thể dự đoán xác suất của tương lai dựa trên dữ liệu lịch sử của

hiện tại. Về bài toán, bối cảnh chuyển trạng thái kết quả của mỗi trận đấu được đánh giá là hợp lý, tuy nhiên, sẽ có một số yếu tố cũng đáng lưu ý, ví dụ như trạng thái thi đấu từ sân nhà sang sân khách và ngược lại. Trong tương lai, em sẽ mở rộng nghiên cứu với vấn đề này.

V. KẾT LUẬN

Như vậy, đồ án này đã báo cáo thực hiện thí nghiệm với ba bài toán về xác suất thống kê ứng dụng, vận dụng các kiến thức được tiếp cận trong môn học để giải quyết các vấn đề. Kết quả thu được ở các bài toán đã cho thấy được tầm quan trọng của xác suất thống kê nói riêng và toán ứng dụng nói chung trong nhiều vấn đề của đời sống, đặc biệt là trong bóng đá, một lĩnh vực mà nhu cầu về thống kê và xác suất là cực kì cần thiết.

Mặc dù trong việc thực hiện mỗi bài toán, việc phát sinh các sai sót dù là chủ quan hay khách quan cũng đều không thể tránh khỏi, nhưng việc thực hiện các bài toán cũng đã giúp cho em (người thực hiện đồ án) hiểu rõ bản chất và các vấn đề đang gặp phải, từ đó có thể mở rộng đề tài và đề xuất các hướng đi trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Konstantinos Loukas, Dimitrios Karapiperis, Georgios Feretzakis, Vassilios S. Verykios, “Predicting Football Match Results Using a Poisson Regression Model,” *Appl. Sci.*, 2024, doi: 10.3390/app14167230.

[2] Germán Coloma, “The Penalty-Kick Game Under Incomplete Information,” *SSRN Electron. J.*, 2012, doi: 10.2139/ssrn.2117476.

[3] Franka Miriam Bruckler, “Probability, statistics and football,” 2015.

[4] Alex Rathke, “An examination of expected goals and shot efficiency in soccer,” *J. Hum. Sport Exerc.*, vol. 12, no., 2017, doi: 10.14198/jhse.2017.12.Proc2.05.

[5] Germán Coloma, “Penalty Kicks in Soccer: An Alternative Methodology for Testing Mixed-Strategy Equilibria,” *J. Sports Econ. - J SPORT ECON*, vol. 8, pp. 530–545, 2007, doi: 10.1177/1527002506289648.

[6] Pablo L. Landeros, “World Cup Penalty Shootouts.”. https://www.kaggle.com/datasets/pablollanderos33/world-cup-penalty-shootouts

[7] Michael Baron, “Probability and Statistics for Computer Scientists; Second Edition,” 2006.

[8] Mike Hughes, Julia Wells, “Analysis of penalties taken in shoot-outs,” *Int. J. Perform. Anal. Sport*, vol. 2, pp. 55–72, 2002, doi: 10.1080/24748668.2002.11868261.

[9] Dương Tôn Đảm, Dương Tôn Thái Dương, *Xác suất thống kê chuyên sâu*. Nhà xuất bản Đại học Quốc gia Thành phố Hồ Chi Minh, 2022.

[10] Alan Agresti, “Introduction to Categorical Data Analysis”, 2018.

[11] B S Everitt, A Skrondal, “The Cambridge Dictionary of Statistics”.

[12] Amitav Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar, S. Chaudhury, “Hypothesis testing, type I and type II errors”, *Ind. Psychiatry J.*, vol. 18, no. 2, pp. 127–131, 2009, doi: 10.4103/0972-6748.62274.

[13] Bruce Bukiet, Elliotte Rusty Harold, José Luis Palacios, “A Markov Chain Approach to Baseball”, *Oper. Res.*, vol. 45, no. 1, 1997, doi: 10.1287/opre.45.1.14.

[14] Supratim Haldar, “Predicting T20 Cricket Results with Markov Chains”, 2017. https://supratim-haldar.medium.com/predicting-t20-cricket-results-with-markov-chains-a-beginners-guide-5b40014e125a

[15] Harvey Campos-Chavez, Soren Thrawl, Anthony DeLegge, A. Harsy, “Predictive Modeling and Analysis of Hockey Using Markov Chains”, vol. 4, no. 1, 2022.

[16] Mert Bayraktar, “English Premier League Matches 2023/2024 Season.” https://www.kaggle.com/datasets/mertbayraktar/english-premier-league-matches-20232024-season

[17] Sheldon M. Ross, *Introduction to Probability Models; Thirteenth Edition*. 2023.

[18] A. Satoh, *Introduction to Molecular-Microsimulation for Colloidal Dispersions*. Elsevier, 2003.

[19] Đoàn Gia Dũng, Hà Thị Phương Thảo, “Một vài ứng dụng của Chuỗi Markov”, 2015.