ỨNG DỤNG PHƯƠNG PHÁP HỒI QUY TỪNG ĐOẠN NÂNG CAO TRONG BÀI TOÁN CHĂM SÓC KHÁCH HÀNG CỦA NGÂN HÀNG

ThS. NGUYỄN VĂN SƠN
Khoa Cơ bản, Học viện Kỹ thuật mật mã
ThS. VŨ DUY HIẾN
Khoa Hệ thống thông tin quản lý, Học viện Ngân hàng
ThS. NGUYỄN VĂN TRUNG
Trung tâm CNTT, Học viện Ngân hàng

Trong những năm gần đây, nhiều nhà khoa học đã nghiên cứu, mô hình hóa các bài toán chuỗi thời gian thực tế trong lĩnh vực tài chính, ngân hàng và ứng dụng các kỹ thuật học máy để giải quyết chúng. Trong đó, kỹ thuật hồi quy tuyến tính được sử dụng phổ biến bởi tính đơn giản, dễ dàng cài đặt và thời gian thực thi ngắn. Tuy nhiên, giả định về mối quan hệ tuyến tính có thể hạn chế các ứng dụng của hồi quy vì nhiều vấn đề kinh doanh và kinh tế là phi tuyến tính về bản chất. Để cải thiện chất lượng của các mô hình hồi quy tuyến tính, chúng tôi đề xuất thuật toán chia khoảng dữ liệu phù hợp để có thể áp dụng kỹ thuật hồi quy tuyến tính từng đoạn. Thuật toán phân chia dựa vào kiểm định phân phối nhằm đảm bảo dữ liệu trong các khoảng chia không bị mất đặc tính phân phối được giả định ban đầu. Các thí nghiệm được thực hiện trên hai bộ dữ liệu bao gồm một bộ dữ liệu tự sinh và một bộ dữ liệu thực mô tả số lượng cuộc gọi đến trung tâm chăm sóc khách hàng của một ngân hàng ở Israel. Kết quả thực nghiệm chỉ ra rằng, độ lỗi của ước lượng sử dụng hàm tuyến tính từng đoạn nhỏ hơn ước lượng sử dụng hàm tuyến tính với mức độ cải thiện đáng kể. Bên cạnh đó, mô hình trong kết quả thực nghiệm có thể được sử dụng để dự đoán số lượng cuộc gọi tới trung tâm chăm sóc khách hàng của các ngân hàng phục vụ cho việc chuẩn bị nguồn lực phù hợp.

I. Giới thiệu

Hoạt động của ngành Tài chính - Ngân hàng đóng một vai trò quan trọng trong việc thiết lập sự ổn định tài chính của mỗi quốc gia. Hơn nữa, toàn cầu hóa và tiến bộ công nghệ đã tạo ra một thị trường cạnh tranh cao cho các ngân hàng. Do đó, những người ra quyết định trong ngành này rất cần các công cụ phân tích dữ liệu lớn, dự đoán, dự báo thông tin để có thể đưa ra quyết định chính xác. Trong những năm gần đây, nhiều nhà nghiên cứu đã mô hình hóa các bài toán chuỗi thời gian thực tế trong lĩnh vực tài chính, ngân hàng và ứng dụng các kỹ thuật học máy để giải quyết chúng. Điển hình là Tanaka et al. (2016), Alessi và Detken (2018) đã sử dụng thuật toán Rừng Ngẫu Nhiên (Random Forest) để cải thiện chất lượng của mô hình cảnh báo sớm khả năng phá sản của ngân hàng. Slavici và cộng sự (2016) đã sử dụng mạng nơ-ron nhân tạo (Artificial Neural Network) để dự báo tình trạng khó khăn tài chính ở Đông Âu. Inam và cộng sự (2018) đã so sánh các kỹ thuật phân tích đa biệt thức (Multivariate Discriminant Analysis), hồi quy Logarit (Logarithmic Regression) và mạng nơ-ron nhân tạo cho bài toán dự đoán phá sản. Tuy nhiên, các kỹ thuật này vẫn phải đối mặt với nhiều thách thức trong thực tế như sự phụ thuộc vào tính sẵn sàng và chất lượng dữ liệu, cài đặt phức tạp và thời gian thực thi cao.

Trong nhiều tài liệu nghiên cứu khác, một kỹ thuật thống kê cũng được sử dụng phổ biến để dự đoán thông tin liên quan đến hoạt động của ngân hàng là

hồi quy tuyến tính. Các nhà kinh tế và nhà phân tích kinh doanh từ lâu đã giả định các mối quan hệ tuyến tính giữa nhiều biến số kinh doanh và kinh tế. Ví dụ đơn giản được thấy trong kinh tế học Keynes (Blinder, 2021), nơi tiêu dùng được biểu thị dưới dạng một hàm tuyến tính của thu nhập. Trong tài chính, mô hình định giá tài sản vốn (CAPM) thể hiện lợi tức kỳ vọng đối với bất kỳ tài sản đảm bảo nào dưới dạng hàm tuyến tính của lợi tức thị trường vượt quá tài sản phi rủi ro. Hơn nữa, trong kỹ thuật hồi quy tài chính quốc tế cũng có thể được sử dụng để đánh giá mức độ ảnh hưởng kinh tế bằng cách phân tích các luồng tiền trong lịch sử và dữ liệu tỷ giá hối đoái (Madura, 2015). Giả định về mối quan hệ tuyến tính có thể hạn chế các ứng dụng của hồi quy vì nhiều vấn đề kinh doanh và kinh tế là phi tuyến tính (về mặt bản chất). Trong những trường hợp như vậy, một số dạng mô hình phi tuyến tính hoặc đường cong tuyến tính có thể được áp dụng. Tuy nhiên, hồi quy tuyến tính được cho là rất hữu ích trong nhiều trường hợp, ví dụ như việc xác định xu hướng dữ liệu bởi tính đơn giản, dễ cài đặt và thời gian thực thi ngắn (Abu Bakar và cộng sự, 2009). Wu và Li (2017) đã chỉ ra rằng, các nhà quản lý gặp rất nhiều khó khăn trong việc áp dụng các mô hình phi tuyến tính vì sự phức tạp của chúng như so với mô hình tuyến tính. Để đơn giản hóa vấn đề này, nhiều nghiên cứu đã thực hiện chuyển đổi một số mô hình phi tuyến tính thành các dạng tuyến tính gần đúng. Phép biến đổi logarit của mô hình phi tuyến tính thành dạng tuyến tính không phải là một ý tưởng mới lạ trong các tài liệu nghiên cứu, các quy trình và cơ chế chuyển đổi, được minh họa bởi Benoit (2011), Rusov và cộng sự (2017), Ogwang (2021). Tuy nhiên, đối với những vấn đề đã được chỉ ra là phù hợp với mô hình phi tuyến tính, việc áp dụng các kỹ thuật chuyển đổi thành dạng tuyến tính có thể làm mất đi một hoặc một vài đặc tính của dữ liệu. Do đó, để có thể đạt được một ước lượng tốt mà không làm tăng độ phức tạp của mô hình, một số nhà khoa học đã sử dụng các hàm tuyến tính từng đoạn (Piecewise-linear fuction) như Brown và cộng sự (2005), Alizadeh và cộng sự (2008). Bên cạnh đó, các mô hình này đều giả định dữ liệu thu thập được có phân phối đã biết. Về mặt lý thuyết, một số quá trình ngẫu nhiên trong thực tế đã được chứng minh có đặc tính của phân phối cụ thể. Ví dụ, số lượng cuộc gọi đến trung tâm chăm sóc khách hàng của một ngân hàng tại một thời điểm tuân theo phân phối Poisson, tỷ giá đồng tiền của một quốc gia (khác với đô-la Mỹ) so với đồng đô-la Mỹ tuân theo phân phối chuẩn. Tuy nhiên, trong thực tế, do quá trình thu thập dữ liệu hoặc cách chia khoảng dữ liệu để ước lượng hàm tuyến tính từng đoạn, dữ liệu thu được không thực sự có tính chất của phân phối được giả định. Ví dụ, phân phối của dữ liệu số lượng cuộc gọi đến trung tâm chăm sóc khách hàng của một ngân hàng tại một thời điểm phụ thuộc vào cách làm tròn đơn vị thời gian (làm tròn đến phút, giờ,...).

Trong bài viết này, đóng góp chính của chúng tôi như sau:

(i) Đề xuất một quy trình học tập đối với dữ liệu chuỗi thời gian hữu hạn cho bài toán hồi quy tuyến tính từng đoạn; (ii) Đề xuất một thuật toán phân chia dữ liệu. Trong thuật toán này, phân phối giả định sẽ được kiểm định trên từng khoảng dữ liệu nhằm đảm bảo việc chia khoảng dữ liệu không làm mất đi đặc tính phân phối của dữ liệu; (iii) Thực nghiệm dựa trên 02 bộ dữ liệu bao gồm: Dữ liệu ngẫu nhiên sinh bởi một hàm tuyến tính từng đoạn cho trước và dữ liệu số lượng cuộc gọi của khách hàng tới trung tâm chăm sóc khách hàng của một ngân hàng ở Israel. Thí nghiệm chỉ ra rằng, thuật toán hồi quy sử dụng hàm tuyến tính từng đoạn có thể thu được kết quả tốt đối với dữ liệu của chúng tôi. Đồng thời, kết quả mô hình có thể được sử dụng để dự đoán số lượng cuộc gọi tới trung tâm chăm sóc khách hàng của ngân hàng phục vụ cho việc chuẩn bị nguồn lực phù hợp.

II. Phát biểu bài toán

Cho G là bộ sinh véc-tơ x trong một quá trình chuỗi thời gian, trong đó $x \in X \subset R^d$ là các quan sát độc lập có cùng phân phối chưa biết (nhưng cố định). Mỗi véc-tơ x có thời điểm quan sát là $t(a \leq t \leq b)$. Một máy học (Learning Machine) quan sát các cặp:

$(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)$

Trong đó, $\{x_i\}_1^n$ là các véc-tơ đầu vào có thời điểm xuất hiện là t_i và $\{y_i\}_1^n$ là phản hồi của người giám sát. Giả sử các véc-tơ đầu vào xuất hiện ngẫu nhiên và độc lập theo phân phối P(x). Từ đó, phản hồi của người giám sát nhận được ngẫu nhiên từ phân phối có điều kiện P(y|x). Trong trường hợp này, tồn tại một phân phối xác suất đồng thời P(x,y) là một phân phối xác suất chưa biết. Mục đích chính của máy học là dự đoán ra một giá trị gần đúng phản hồi của người giám sát y_i trên bất kỳ vectơ đầu vào x_i nào được tạo bởi bộ sinh *G*. Hàm xấp xỉ được chọn từ một không gian hàm giả thuyết *F* cho trước:

$$F = \{f(x,\alpha) | f \in L_2(P), \alpha \in \Lambda\}$$

Để chọn hàm hồi quy tốt nhất, chúng ta cần tối thiểu sự mất mát hoặc sự khác biệt giữa phản hồi của người giám sát và phản hồi của máy học đối với một đầu vào nhất định thông qua hàm rủi ro: $R = \int (y-f(x,\alpha))^2 \, dP(x,y)$

Theo Vapnik (1995), bài toán tối thiểu hàm rủi ro có thể được quy về bài toán tối thiểu hàm rủi ro thực nghiệm:

$$R_{emp} = \frac{1}{n}\sum_{1}^{n}(y_i - f(x_i, \alpha))^2$$

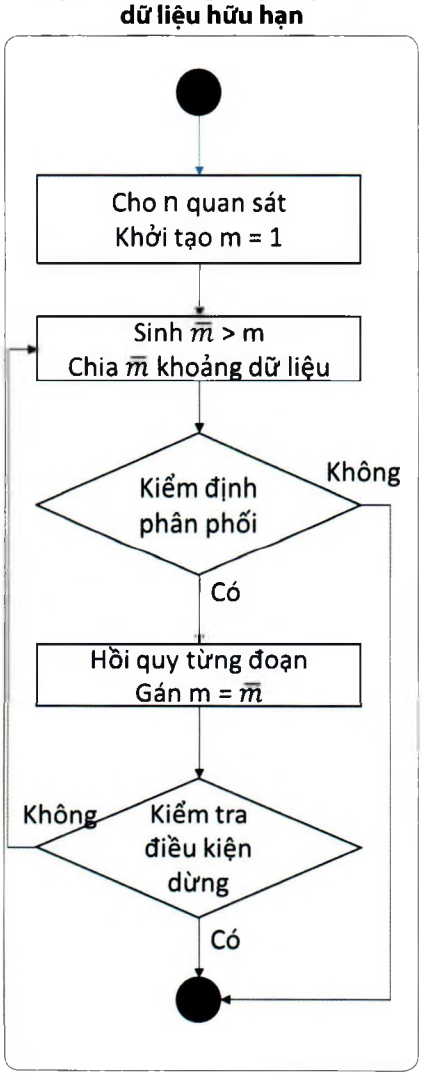
Giả sử không gian hàm giả thuyết được cố định là không gian các hàm tuyến tính nhằm làm giảm độ phức tạp của mô hình (nghĩa là thuộc tính số chiều của bộ dữ liệu đạt được giá trị nhỏ hơn, theo Vapnik, 1995). Nếu dữ liệu thu thập được có xu hướng của hàm phi tuyến, bài toán đặt ra là làm sao giảm độ lỗi của mô hình học khi không gian hàm giả thuyết đã được cố định. Trong bài viết này, chúng tôi đề xuất phương pháp ước lượng trên từng đoạn, cụ thể là, hàm tuyến tính trên từng đoạn dựa trên dữ liệu chia khoảng theo thời gian quan sát. Đối với mỗi khoảng dữ liệu, chúng tôi thực hiện các phương pháp kiểm định phân phối nhằm đảm bảo dữ liệu sau khi chia khoảng vẫn tuân theo phân phối như giả thiết ban đầu.

III. Thuật toán

Chúng tôi giả sử rằng dữ liệu huấn luyện được lưu trữ trong khoảng không chồng lên nhau $P_k=[u_{k-1},u_k], k=\overline{1,m}$ được chia bởi các điểm chia $a=u_0,u_1,u_2,...,u_m=b$. Với một khoảng chia nhỏ, một hàm tuyến tính có thể dễ dàng khớp với dữ liệu huấn luyện nhưng cũng làm tăng xác suất xảy ra trường hợp "overfitting". Với khoảng chia lớn, mô hình có thể bị "underfitting" do ta cố gắng mô tả các dữ liệu phức tạp bằng các mô hình tuyến tính đơn giản. Do đó, với dữ liệu hữu hạn, một thuật toán phân chia tốt cần đảm bảo tối đa số điểm chia và dữ liệu trong mỗi khoảng chia phải tuân theo cùng một phân phối giả định. Việc đảm bảo đặc tính phân phối dữ liệu nhằm tăng khả năng khái quát hóa của mô hình với những dữ liệu chưa thu thập được. Hình 1 mô tả quy trình đề xuất các hàm tuyến tính từng đoạn dựa trên dữ liệu hữu hạn.

Việc phân chia các khoảng dữ liệu được thực hiện liên tục nếu các khoảng

Hình 1: Quá trình hồi quy tuyến tính từng đoạn dựa trên dữ liệu hữu hạn



Thuật toán 1: Thuật toán hồi quy tuyến tính từng đoạn với phương pháp chia khoảng dữ liệu có kiểm định phân phối

Bước 1: Khởi tạo: $er = \infty$; $D = \{a, b\}$; i = 0;

Bước 2: Gán $u = a$; $v = b$;

Bước 3: Nếu $i > maxIterations$:

Đi đến Bước 7;

Bước 4: Chọn ngẫu nhiên t ∈ (u, v);

Bước 5: Phân chia dữ liệu:

$S_1 = \{x_i | x_i \in [u, t]\}$;

$S_2 = \{x_i | x_i \in [t, v]\}$;

Bước 6: Kiểm định phân phối cho 02 bộ dữ liệu thuộc tập S_1, S_2:

Nếu S_1, S_2 tuân theo phân phối giả định:

e_1 = Độ lỗi của ước lượng hàm tuyến tính trên đoạn [u, t];

e_2 = Độ lỗi của ước lượng hàm tuyến tính trên đoạn [t, v];

Nếu $e_1 + e_2 < er$:

$D = D \cup \{t\}$;

Quay lại Bước 2 với er = $e_1 + e_2$, u = u, v = t;

Quay lại Bước 2 với er = $e_1 + e_2$, u = t, v = v;

Ngược lại

i++;

Quay lại Bước 2;

Ngược lại:

i++;

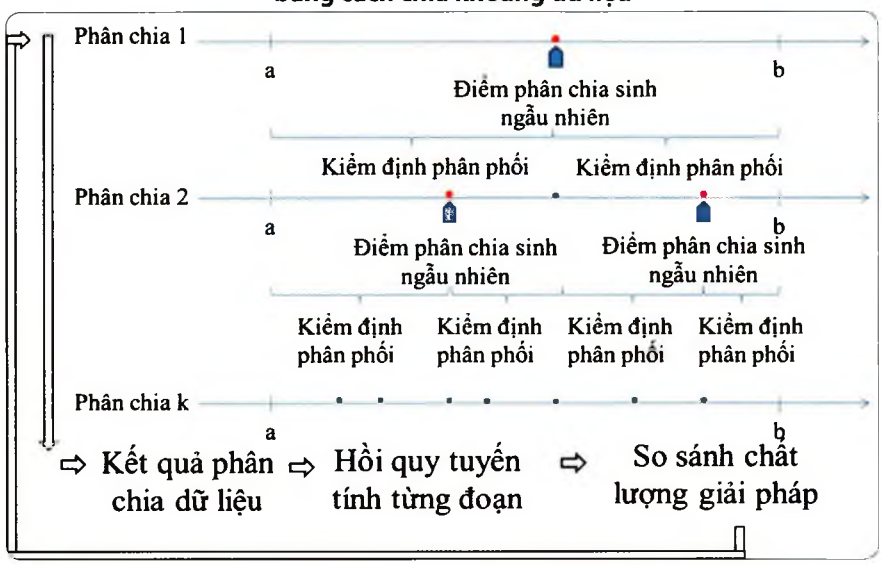
Quay lại Bước 2;

Bước 7: Sắp xếp lại các điểm chia thuộc D theo thứ tự tăng dần;

Bước 8: Ước lượng các hàm tuyến tính từng đoạn với các điểm chia thuộc D;

dữ liệu vẫn còn tuân theo phân phối giả định hoặc lỗi trên tập dữ liệu xác thực không tăng (chưa xảy ra). Chi tiết thuật toán được mô tả trong Thuật toán 1. Trong Thuật toán 1, danh sách các điểm phân chia được lưu trong tập *D*. Tại mỗi lần lặp, một điểm chia được sinh ngẫu nhiên trong khoảng *(u,v)*. Nếu điểm chia này tạo thành hai tập dữ liệu tuân theo phân phối giả định thì tiếp tục quay lại Bước 2 với các khoảng xem xét mới. Ở bước này, chúng ta có thể sử dụng một trong các phép kiểm định phân phối phổ biến như Kolmogorov-Smirnov Test hoặc Log-Test. Thủ tục lặp lại cho đến khi gặp giới hạn số lần thực thi. Quá trình thực thi của thuật toán này được mô tả trực quan trong Hình 2.

Hình 2: Mô tả thuật toán hồi quy tuyến tính từng đoạn bằng cách chia khoảng dữ liệu



IV. Thực nghiệm

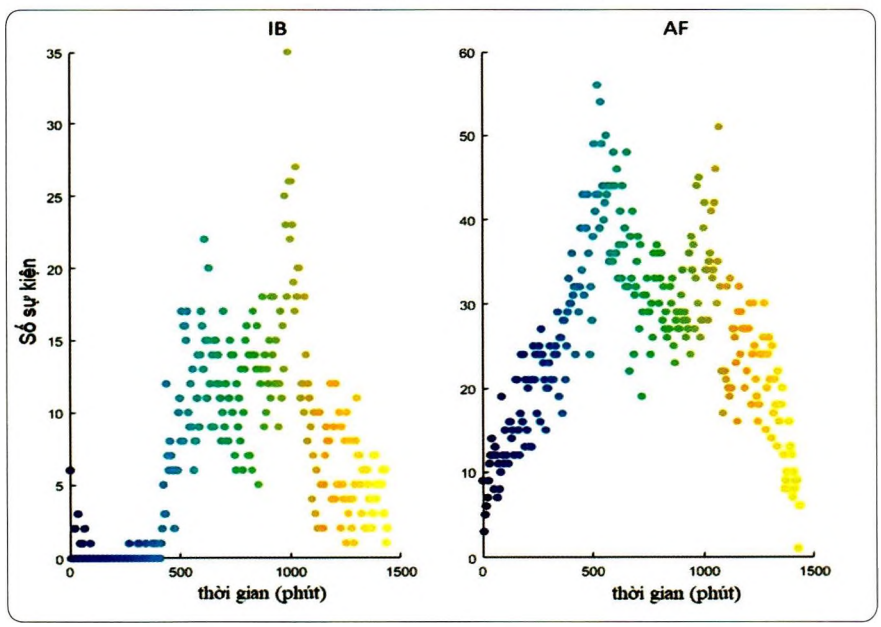
Trong bài viết này, chúng tôi sử dụng 02 bộ dữ liệu để thử nghiệm được mô tả như sau:

- Dữ liệu ngẫu nhiên sinh bởi hàm tuyến tính từng đoạn:

$$f^*(t)=\begin{cases}\frac{13}{36}t+7, & 0\le t<10800\\ \frac{5}{36}t+15, & 10800\le t<21600\\ \frac{25}{36}t-25, & 21600\le t<32400\\ \frac{-1}{2}t+104, & 32400\le t<43200\\ \frac{-1}{18}t+40, & 43200\le t<54000\\ \frac{1}{3}t-30, & 54000\le t<64800\\ \frac{-4}{9}t+138, & 64800\le t<75600\\ \frac{-5}{9}t+166, & 75600\le t<86400\\ 0, & otherwise\end{cases}$$

Trong đó, t là thời gian sự kiện xuất hiện (đơn vị: giây), *f*(t)* là hàm mô tả số lượng sự kiện xảy ra tại thời điểm t (ký hiệu là AF).

Hình 3: Trực quan hóa dữ liệu



- Dữ liệu số lượng cuộc gọi thực tế đến trung tâm chăm sóc khách hàng của một ngân hàng ở Israel mỗi ngày trong khoảng thời gian 12 tháng năm 2020 (ký hiệu là IB).

Cả hai bộ dữ liệu này được giả định tuân theo phân phối Poisson.

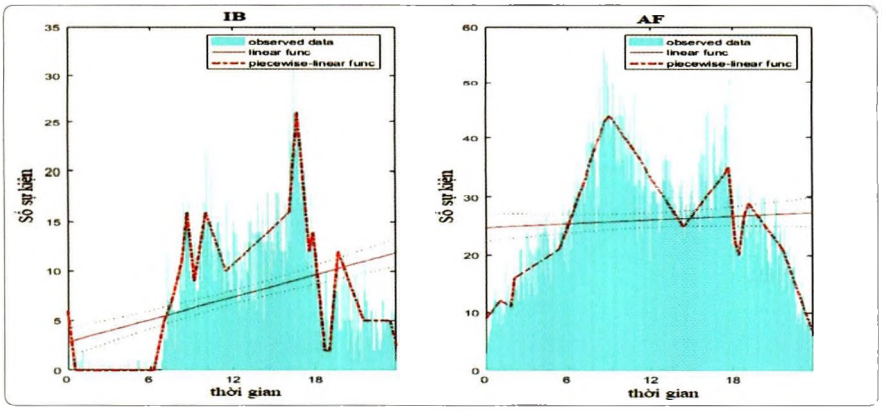
1. Trực quan hóa dữ liệu

Đầu tiên, chúng tôi vẽ các biểu đồ phân tán cho các điểm dữ liệu để khảo sát sự tồn tại của tính chất tuyến tính. Sự tồn tại của tuyến tính là các điểm được phân bố đối xứng xung quanh một đường chéo. Hình 3 hiển thị 02 biểu đồ tương ứng mô tả phân bố tập dữ liệu IB (trái) và AF (phải). Chúng ta có thể thấy với tập AF, tính chất tuyến tính từng đoạn được thể hiện khá rõ ràng. Đối với tập IB, dữ liệu phân tán rộng nhưng vẫn có yếu tố đối xứng quanh một đường chéo trên từng đoạn.

2. Ước lượng hàm tuyến tính từng đoạn

Bảng 1 hiển thị kết quả ước tính theo cả hàm tuyến tính và hàm tuyến tính từng đoạn. Kết quả được trực quan hóa trong Hình 4. Trong Bảng 1, er_1 và er_2 tương ứng biểu thị giá trị trung bình của độ lỗi bình phương (trên tập kiểm tra)

Hình 4: Kết quả ước lượng hàm Rate-function của 02 bộ dữ liệu thử nghiệm



Bảng 1: Kết quả quá trình hồi quy

Bộ dữ liệu	Thuật toán Hồi quy tuyến tính		Thuật toán Hồi quy tuyến tính từng đoạn			Mức độ cải thiện lỗi
	er_1	$t_1(s)$	er_2	$t_2(s)$	Số đoạn	ρ
IB	6.06	47	2.73	267	21	54.95
AF	10.6	35	4.74	188	16	55.28

ước lượng bằng phương pháp hồi quy tuyến tính thông thường và phương pháp hồi quy tuyến tính từng đoạn được đề xuất. Mức độ cải thiện của mô hình so với ước lượng bằng hàm tuyến tính thông thường được tính theo công thức: $\rho = \frac{er_1 - er_2}{er_1}$

Với số vòng lặp tối đa maxIterations =10000, chúng tôi nhận thấy rằng độ lỗi của ước lượng theo hàm tuyến tính từng đoạn nhỏ hơn độ lỗi của ước lượng bằng hàm tuyến tính. Mức độ cải thiện độ lỗi của mô hình đề xuất có giá trị lần lượt đối với hai bộ dữ liệu IB và AF là 54,95% và 55,28%. Hơn nữa, số khoảng dữ liệu chia được tương ứng với hai bộ dữ liệu là 21 và 16. Điều này cho thấy, khoảng thời gian đang xem xét có thể phân chia thành các khoảng con mà không mất đi đặc tính phân phối của dữ liệu. Từ Hình 4, chúng ta dễ dàng nhận thấy mô hình hồi quy sử dụng hàm tuyến tính từng đoạn có độ khớp tốt với xu hướng của dữ liệu.

Từ kết quả trên cho thấy, phương pháp ước lượng sử dụng hàm tuyến tính từng đoạn cho kết quả xấp xỉ tốt hơn hàm tuyến tính. Đồng thời, việc sử dụng các bài kiểm định phân phối đảm bảo việc phân chia khoảng không làm mất đi thuộc tính phân phối của dữ liệu. Từ đó tăng khả năng khái quát hóa cho mô hình học.

V. Kết luận

Trong bài viết này, chúng tôi đã đề xuất thuật toán hồi quy tuyến tính từng đoạn thay thế thuật toán hồi quy tuyến tính thông thường. Các kết quả thực nghiệm đã cho thấy mức độ cải thiện lỗi ρ của thuật toán đề xuất so với thuật toán truyền thống đã được tăng lên đáng kể (trên 50%). Đồng thời, để chứng minh khả năng ứng dụng của kết quả nghiên cứu, chúng tôi đã ứng dụng thuật toán đề xuất vào bài toán thực tế dự đoán số lượng cuộc gọi tới trung tâm chăm sóc khách hàng của ngân hàng và đã cho kết quả rất khả quan. Trong thời gian tới, chúng tôi sẽ tập trung cải thiện hơn nữa chất lượng của mô hình dự báo dựa trên phương pháp hồi quy tuyến tính.■

TÀI LIỆU THAM KHẢO:

1. Abu Bakar, Nor Mazlina & Mohd Tahir, Izah. (2009). Applying Multiple Linear Regression and Neural Network to Predict Bank Performance. International Business Research. 2. 10.5539/ibr.v2n4p176.

2. K. Tanaka, T. Kinkyo, S. Hamori Random forests-based early warning system for bank failures Econ. Lett., 148 (2016), pp. 118-121.

3. L. Alessi, C. Detken Identifying excessive credit growth and leverage J. Financ. Stabil., 35 (2018), pp. 215-225

4. Geng, R., Bose, I. and Chen, X. (2015), "Prediction of financial distress: an empirical study of listed Chinese companies using data mining", European Journal of Operational Research, Vol. 241 No. 1, pp. 236-247.

5. Slavici, T., Marris, S. and Pirtea, M. (2016), "Usage of artificial neural networks for optimal bankruptcy forecasting Case study: Eastern european small manufacturing enterprises", Quality and Quantity, Vol. 50 No. 1, pp. 385-398.

6. Inam, F., Inam, A., Mian, M.A., Sheikh, A.A. and Awan, H.M. (2018), "Forecasting bankruptcy for organizational sustainability in Pakistan: using artificial neural networks, logit regression, and discriminant analysis", Journal of Economic and Administrative Sciences, Vol. 35 No. 3, pp. 183-201.

7. Blinder, Alan S. "Keynesian Economics". www.econlib.org. The Concise Encyclopedia of Economics. Retrieved 13 March 2021.

8. Madura, J (2015), "International Financial Management" 12th Edition, Cengage Learning, New Tech Park, Singapore.

9. Ogwang, John. "Some Non-Linear Problems in Accounting and Finance: Can We Apply Regression?" International journal of business and economics 8 (2021): 81-99.

10. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical analysis of a telephone call center: a queueing-science perspective. J Amer Stat Assoc 100:36-50.

11. Alizadeh, F., Eckstein, J., Noyan, N., \& Rudolf, G. (2008). Arrival rate approximation by nonnegative cubic splines. Operations Research, 56(1), 140-156.

12. V. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.