PHÁT HIỆN MẪU BẤT THƯỜNG CHO TRONG DOANH NGHIỆP BÁN LẺ BẰNG PHÂN TÍCH MOTIF

FRAUD PATTERN DETECTION BY MOTIF DISCOVERY IN RETAIL BUSINESS

Phạm Ngọc Quang Anh[1,2], Vũ Thành Nam[1], Hoàng Văn Đông[1], Lê Anh Ngọc[3], Nguyễn Thị Ngọc Anh[1]

1. Viện Toán ứng dụng và Tin học, Đại học Bách khoa Hà Nội
2. Viện nghiên cứu ứng dụng Công nghệ CMC
3. Đại học FPT, Việt Nam

Ngày nhận bài: 12/11/2021, Ngày chấp nhận đăng: 06/06/2022, Phản biện: TS Nguyễn Thị Thanh Tân

Tóm tắt:

Những khách hàng xấu thực hiện các hành vi gian lận trong các giao dịch tài chính gây thiệt hại về kinh tế và là mối nguy hiểm cho các công ty, tổ chức. Trong thời gian gần đây, các giao dịch bùng nổ bởi sự phát triển của các giao dịch tài chính qua mạng và di động trên toàn thế giới. Vì vậy, việc xử lý giao dịch và phát hiện những hành vi bất thường từ hàng trăm ngàn giao dịch với vô số loại hành vi khác nhau không còn phù hợp với phương thức xử lý thủ công. Trong bài báo này việc khai phá motif cho chuỗi thời gian và phát hiện bất thường bằng thuật toán học máy rừng ngẫu nhiên được đề xuất. Một mô hình xác định các mẫu hành vi gian lận và phân loại các đối tượng trong bài toán phát hiện bất thường ở cấp độ tài khoản được mô hình hoá. Mô hình đề xuất được thử nghiệm và sau đó sử dụng để phát hiện ra các khách hàng bất thường trong dữ liệu hoạt động bán lẻ. Bằng thực nghiệm chỉ ra mô hình có độ chính xác F1 là 75%.

Từ khóa:

Phát hiện bất thường, khai phá mô-típ, nhận dạng mẫu, học máy

Abstract:

Bad customers do fraud behavior in financial transactions is cause of economic losses and the threat for companies and organizations. Transactions exploded recently because of the development of mobile, online transactions in whole the world. Therefore, transaction processing and detecting anomalous behavior from hundreds of thousands of transactions with various types of behavior are no longer suitable for manual processing. In this paper, motif discovery for time series and anomaly detection by machine learning are proposed. A model that identifies fraudulent behavior patterns and classifies objects in the account level anomaly detection problem is modeled. The proposed model is experimented and then used to discover anomalies in retail activity data. Experimentally, the model has an accuracy of F1 of 75%

Keywords:

Fraud detection, motif discovery, pattern recognition, machine learning

1. Giới thiệu chung

Bài toán phát hiện gian lận là một chủ đề quan trọng của các công ty, tổ chức như ngân hàng, bảo hiểm, các doanh nghiệp bán lẻ [1]. Những gian lận trong tài chính sẽ ảnh hưởng đến uy tín, gây tổn thất cho các tổ chức này. Bài toán này được rất

nhiều nhà nghiên cứu quan tâm và thực hiện hàng loạt công trình nghiên cứu trong hai thập kỉ gần đây [1],[2],[3],[4],[5].

Các giao dịch tài chính bùng nổ gần đây bởi sự phát triển của các giao dịch tài chính qua mạng và di động. Các hình thức giao dịch này có chi phí thấp và rất nhanh chóng, tiện lợi. Từ sự phát triển này, việc xử lý giao dịch từ hàng trăm ngàn giao dịch với vô số loại hành vi khác nhau không còn phù hợp với phương thức xử lý thủ công. Nhiều ứng dụng của phân tích dữ liệu và trí tuệ nhân tạo được áp dụng để giải quyết vấn đề này như phân lớp, phân nhóm, hồi quy, phát hiện ngoại lai, dự đoán và mô phỏng [1].

Việc tiếp cận bài toán phát hiện gian lận có thể chia vào 3 mức độ: cấp độ giao dịch, cấp độ tài khoản, cấp độ mạng. Thứ nhất, cấp độ giao dịch tiếp cận chủ yếu trên các giao dịch và các thuộc tính. Thứ hai, cấp độ tài khoản chủ yếu trên các đặc điểm thống kê và các hành vi của tài khoản (người, công ty, tổ chức) trong các giao dịch. Thứ ba, cấp độ mạng, một mạng đồ thị thể hiện mối quan hệ giữa các tài khoản của mỗi giao dịch.

Khi các chuyên gia phân tích đánh giá rủi ro một đối tượng nào đó, một trong những cách họ thường sử dụng là xem xét quá trình hoạt động của đối tượng này trong quá khứ. Họ tổng kết những kinh nghiệm có được từ những đối tượng gian lận, sử dụng những kinh nghiệm này để đối chiếu với đối tượng đang được xét và phân tích xem liệu các hành vi của đối tượng này có giống với những hành vi đáng ngờ hay không.

Quá trình hoạt động của các đối tượng có thể được mô tả bởi những chuỗi thời gian, các hành vi mà đối tượng thực hiện chính là các chuỗi con của chúng. Do đó, muốn thực hiện ý tưởng đã nêu ra, ta phải tìm cách để xây dựng một tập chứa các mẫu hành vi bất thường của những kẻ xấu, đồng thời tạo ra một hệ thống phân loại để sử dụng những mẫu hành vi này vào quá trình đánh giá các đối tượng khác.

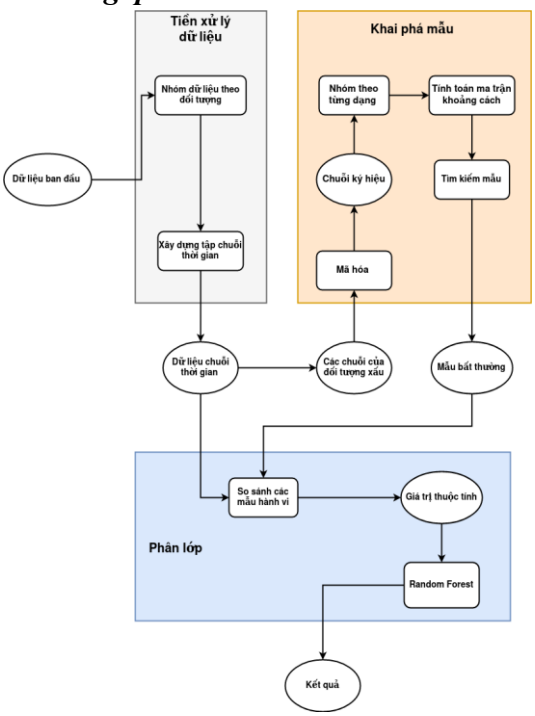
Mục tiêu chính của bài báo này là đề xuất một mô hình mới khai phá các mẫu hành vi, thói quen đáng nghi ngờ mà những đối tượng gian lận thực hiện trong quá trình hoạt động bán lẻ. Từ đó phân lớp đối tượng vào lớp bất thường (có khả năng thực hiện hành vi gian lận) hoặc bình thường. Mô hình được áp dụng trong phạm vi đối tượng khách hàng có hoạt động mua bán nhiều mặt hàng khác nhau, có giao dịch với công ty bán lẻ ở nhiều chi nhánh khác nhau.

Nội dung chính của bài báo được trình bày trong 4 phần. (i) Phần 1 là giới thiệu chung về phát hiện bất thường trong giao dịch. (ii) Phần 2 trình bày phương pháp xây dựng mô hình hệ thống phát hiện bất thường. Các thuật toán tìm kiếm mẫu hành vi đáng ngờ trên chuỗi thời gian, sử dụng thuật toán học máy rừng ngẫu nhiên tiến hành phân lớp đối tượng. (iii) Phần 3 là áp dụng mô hình đưa ra với dữ liệu khách hàng. (iv) Phần 4 là kết

luận và nêu ra hướng nghiên cứu tiếp theo của bài báo.

2. Phương pháp luận

2.1. Tổng quan mô hình đề xuất



Hình 1: Tổng quan mô hình đề xuất

Để tìm được các kịch bản gian lận ẩn trong dữ liệu sẵn có, ta cần xét một chuỗi các giao dịch liên tiếp do cùng một đối tượng thực hiện. Một chuỗi giao dịch như vậy được gọi một hành vi. Như vậy, yêu cầu đặt ra cho hệ thống là tìm ra các mẫu chung của các hành vi do đối tượng gian lận thực hiện trước giao dịch phát sinh gian lận.

Trước hết, tập dữ liệu cần được tiền xử lý bằng cách nhóm theo từng đối tượng và sắp xếp lại theo trình tự thời gian để tạo thành các chuỗi thời gian mô tả quá trình hoạt động của chủ thể giao dịch. Chuỗi thời gian được xây dựng sẽ mô tả một hoặc nhiều mặt của giao dịch (tương ứng với chuỗi thời gian đơn và chuỗi thời gian đa chiều). Lấy ví dụ, trong trường hợp chỉ xét đến sự biến thiên của giá trị giao dịch, ta xây dựng chuỗi thời gian đơn theo thuộc tính tổng giá trị giao dịch theo từng tháng. Mặt khác, nếu muốn thể hiện tổng lợi nhuận và số lượng giao dịch của đối tượng, ta xây dựng chuỗi thời gian hai chiều.

Kết quả của bước tiền xử lý là các hành vi gian lận được thể hiện trực quan trên các chuỗi thời gian. Bài toán phát hiện gian lận ban đầu được chuyển thành tìm các motif lặp lại trên các chuỗi thời gian của những đối tượng xấu.

Trong quá trình tìm kiếm motif cho chuỗi thời gian, các hành vi những đối tượng xấu thường thực hiện và các hành vi đã bị đánh dấu là xấu từ trước sẽ được lọc ra, tổng hợp lại thành tập mẫu hành vi đáng ngờ. Tập mẫu thu được là cơ sở để so sánh, đánh giá những đối tượng khác trong bước tiếp theo.

Cuối cùng, ta sẽ tiến hành phân lớp các đối tượng trong tập dữ liệu thành hai lớp: bất thường hoặc bình thường. Tiêu chuẩn được đặt ra để kết luận nhãn cho đối tượng là mức độ tương đồng trong hành vi của nó với các hành vi nằm trong tập mẫu hành vi đáng ngờ. Do đó, ta phải đưa ra phương thức so sánh độ tương tự giữa các mẫu hành vi ở dạng chuỗi con của chuỗi thời gian với nhau.

Áp dụng độ đo tương tự, quá trình so sánh các mẫu hành vi sẽ cho ta bộ thuộc tính mới của mỗi đối tượng. Trong đó, mỗi thuộc tính sẽ đại diện cho độ tương tự của những hành vi của đối tượng với những hành vi trong tập mẫu đáng nghi. Bộ thuộc tính này sẽ được sử dụng để xác định sự tương đồng giữa đối tượng gian lận và đối tượng được kiểm tra. Sử dụng thuật toán phân lớp, ta tìm ra các doanh nghiệp có nhiều tương đồng với các doanh nghiệp có hành vi gian lận mua

hàng và gán nhãn bất thường cho doanh nghiệp tương ứng.

Quá trình thực hiện cụ thể của các bước kể trên sẽ được thể hiện chi tiết trong các mục tiếp theo.

2.2. Mô hình hóa dữ liệu thành chuỗi thời gian

Một chuỗi thời gian $T = \{x_0, x_1, \dots, x_m\}$ là một dãy thứ tự theo thời gian $\{t_0, t_1, \dots, t_m\}$ của m biến giá trị thực [6].

Trong phạm vi bài báo, ta sẽ chỉ xét loại chuỗi thời gian có tập mốc thời gian cố định. Kí hiệu tập mốc thời gian $\Gamma = \{t_j\}_{0 \le j \le m}$.

Giả sử ta có tập dữ liệu chứa dữ liệu về các giao dịch rời rạc do n đối tượng $x_1, x_2, \dots, x_n$ thuộc tập đối tượng $\mathbb{O}$ thực hiện.

Với mỗi đối tượng x_i trong $\mathbb{O}$, ta nhóm các giao dịch do x_i thực hiện lại và sắp xếp chúng theo thứ tự thời gian. Khi đó, ta có một chuỗi thời gian tổng quát mô tả hoạt động của x_i với các điểm dữ liệu là các vector thuộc tính của những giao dịch do x_i thực hiện.

Trong nhiều trường hợp, chỉ một hoặc một vài thuộc tính trong giao dịch được xét. Ngoài ra, việc phát hiện các đối tượng bất thường cần phải tổng hợp thông tin của giao dịch trong một khoảng thời gian nhất định. Vì vậy, việc xây dựng chuỗi thời gian không dựa trên từng thời điểm giao dịch mà dựa trên một mốc thời gian cố định (chẳng hạn, chuỗi thời gian thể hiện sự thay đổi kê khai mã nhóm hàng hóa của doanh nghiệp theo tháng).

Như vậy, với mỗi đối tượng x_i, ta thành lập một chuỗi thời gian tương ứng với tập mốc thời gian Γ, kí hiệu là TS_i. Ở đó

$$TS^i = \{v_j^i\} \quad j \in \Gamma \quad (1)$$

Trong đó, giá trị của ${v_j}^i$ là một thống kê các thuộc tính đang xét của những giao dịch do đối tượng x_i thực hiện phát sinh trong khoảng thời gian $[t_j, t_{j+1})$. Chuỗi thời gian này thể hiện các hành vi như thay đổi tần suất giao dịch, giá trị giao dịch,...

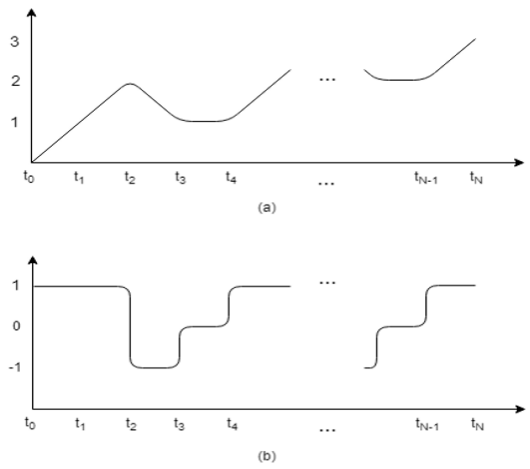
Có nhiều loại chuỗi thời gian với độ phức tạp khác nhau thể hiện mức độ thay đổi của các hành vi. Trong khuôn khổ bài báo, ta xét một chuỗi thời gian đơn giản

Định nghĩa 1. *Một chuỗi thời gian S được gọi là chuỗi thời gian đơn giản nếu thỏa mãn điều kiện sau:*

$$v_{j+1} - v_j \in \{-1, 0, 1\} \quad \forall j \in \{1, 2, \dots, m\} \quad (2)$$

Chuỗi $Z = z_1, z_2, \dots, z_m$ với $z_j = v_{j+1} - v_j$ là phép trừ chuỗi của chuỗi thời gian S.

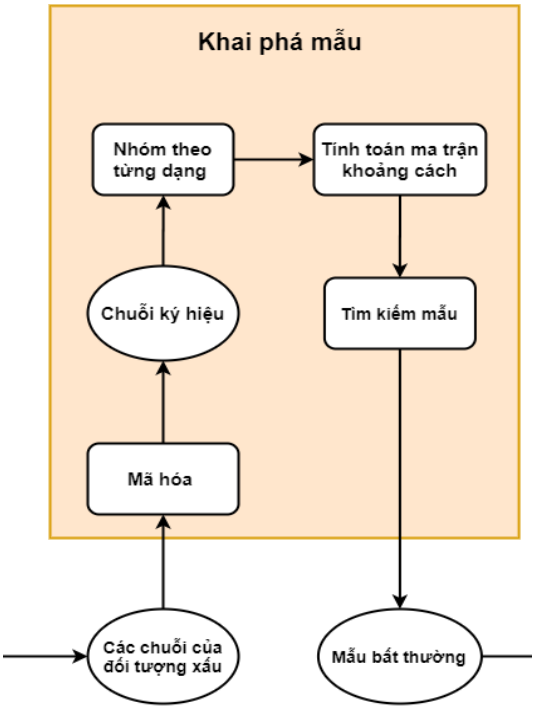
Hình 2 mô tả một chuỗi thời gian đơn giản và phép trừ chuỗi của chuỗi thời gian đó. Ký hiệu TS là tập dữ liệu chuỗi thời gian của tập đối tượng $\mathbb{O}$



Hình 2: Mô tả chuỗi thời gian đơn giản: (a) Chuỗi thời gian. (b) Phép trừ chuỗi của chuỗi thời gian.

2.3. *Khai phá mẫu bất thường*

Chuỗi thời gian được xây dựng trong phần mô hình hóa có thông tin về các hành vi và thói quen của các đối tượng trong quá khứ. Ta tiếp tục phân tích các thông tin này, cụ thể là tìm các mẫu hành vi bất thường của đối tượng gian lận. Hình 3 mô tả lại quy trình khai phá mẫu trong sơ đồ tổng quan.

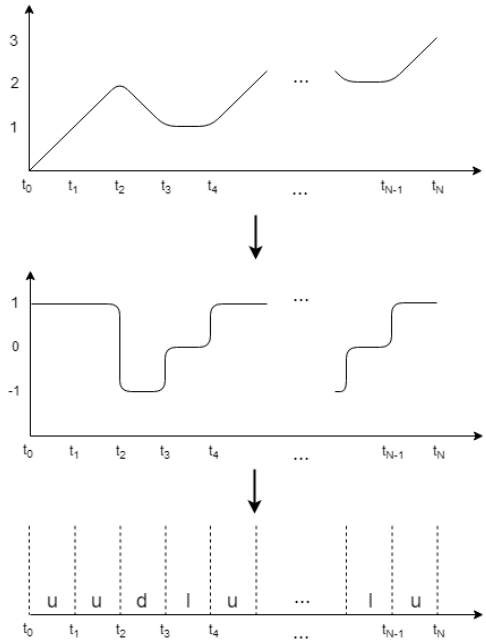


Hình 3 : Quy trình khai phá mẫu.

Xét tập hợp $\mathbb{A} \in TS$ là tập hợp của các chuỗi thời gian tương ứng với đối tượng gian lận. Mục tiêu chính của khai phá mẫu là xây dựng tập chứa các hành động đáng nghi để làm nền tảng cho quá trình phân loại tiếp sau. Ý tưởng được đưa ra ở đây là: hành vi nào do nhiều đối tượng gian lận hay thực hiện thì đáng nghi. Nói cách khác, ta phải phân tích các chuỗi thời gian của các đối tượng xấu để tìm ra những mẫu hành vi xuất hiện nhiều lần.

Nhiều nghiên cứu đã đi theo cách tiếp cận sử dụng biểu diễn rời rạc của chuỗi thời gian. Ý tưởng chính của phương pháp tiếp cận này là mã hoá chuỗi thời gian thành một chuỗi rời rạc đơn giản mà vẫn giữ được thông tin về sự biến động, chẳng hạn như mã hoá thành chuỗi ký tự theo thuật toán SAX [7].

Với mỗi đối tượng x_i , chuỗi Z_i của chuỗi thời gian TS_i là một chuỗi của số thuộc $\{-1,0,1\}$, ta chuyển đổi nó vào chuỗi ký hiệu S với 1 là u , 0 là l và -1 là d. Hình 4 mô tả quá trình chuyển từ chuỗi thời gian thành chuỗi ký hiệu.



Hình 4: Chuyển đổi phép trừ chuỗi của chuỗi thời gian đơn giản và chuỗi ký hiệu.

Định nghĩa 2. *Chuỗi ký hiệu thu gọn là biểu diễn rút gọn cho chuỗi ký hiệu thông thường. Trong chuỗi này, mỗi ký hiệu đi kèm với một chỉ số cho biết rằng ký hiệu đó lặp lại bao nhiêu lần.*

Định nghĩa 3. *Dạng của chuỗi ký hiệu thu gọn là chuỗi các ký hiệu trong chuỗi ký hiệu thu gọn nhưng lược bỏ chỉ số.*

Lấy ví dụ, $u_2 l_4 d_3$ là chuỗi ký hiệu thu gọn của $u-u-l-l-l-l-d-d-d$. Dạng của $u_2 l_4 d_3$ là $u-l-d$.

Trong khuôn khổ bài báo, để đơn giản ta gọi dạng của chuỗi ký hiệu thu gọn là dạng.

Tập chuỗi thời gian TS được chuyển đổi thành tập các chuỗi ký hiệu thu gọn S'. Tương ứng với từng dạng, các chuỗi con của các chuỗi ký hiệu được thu thập và đưa vào một tập chuỗi ký hiệu.

Với mỗi tập, tất cả các chuỗi con được so sánh bởi độ đo khoảng cách sau:

Định nghĩa 4. *Cho hai chuỗi X và X' là hai chuỗi ký hiệu thu gọn với cùng dạng:*

$X = x^1_{a_1} x^2_{a_2} \dots x^q_{a_q}$
$X' = x^1_{a'_1} x^2_{a'_2} \dots x^q_{a'_q}$

Khoảng cách giữa X và X' là:

$$D(X, X') = \sqrt{\frac{(a_1 - a'_1)^2 + (a_2 - a'_2)^2 + \dots + (a_q - a'_q)^2}{q}}$$

Kết quả của các phép so sánh này được thể hiện bởi một ma trận khoảng cách. Ở bước này, ta định nghĩa một ngưỡng tương đồng R, nếu hai chuỗi có khoảng cách nhỏ hơn R thì hai chuỗi tương đồng và có cùng mẫu.

Với mỗi mẫu sẽ có một chuỗi trung tâm của mẫu đại diện cho nó thỏa mãn:

- Là chuỗi có số lượng chuỗi ký hiệu thu gọn lớn nhất mà chuỗi con của chúng tương đồng với chuỗi đang xét.
- Nếu có nhiều hơn một chuỗi thỏa mãn điều kiện trên thì chuỗi trung tâm mẫu là chuỗi có tổng khoảng cách tới tất cả các chuỗi tương đồng với nó là nhỏ nhất.

Nếu thỏa mãn cả hai điều kiện thì ta chọn ngẫu nhiên chuỗi trung tâm mẫu từ các chuỗi đó.

Ví dụ, xét 5 chuỗi ký hiệu sau:

$SSS_1 = u_1 l_1 u_3 l_1 \quad SSS_4 = u_1 l_1 u_2 l_2$
$SSS_2 = u_1 l_1 u_2 l_1 \quad SSS_5 = u_2 l_1 u_1 l_2$
$SSS_3 = u_2 l_1 u_1 l_1$

Sử dụng công thức khoảng cách ở định nghĩa 4

$$D(SSS_i, SSS_j) = \sqrt{\frac{(a_1 - a'_1)^2 + (a_2 - a'_2)^2 + \dots + (a_4 - a'_4)^2}{4}}$$

	SSS_1	SSS_2	SSS_3	SSS_4	SSS_5
SSS_1	0	0.5	1.12	0.71	1.22
SSS_2	0.5	0	0.71	0.5	0.87
SSS_3	1.12	0.71	0	0.87	0.5
SSS_4	0.71	0.5	0.87	0	0.71
SSS_5	1.22	0.87	0.5	0.71	0

Bảng 1: Ma trận khoảng cách

Trong bảng 1, nếu ta chọn ngưỡng tương đồng là 0.75 thì SSS_1, SSS_2, SSS_3 và SSS_4 sẽ có mẫu giống nhau, chỉ SSS_5 là riêng biệt. Chuỗi trung tâm của 4 mẫu tương đồng là SSS_2.

Có thể có nhiều hơn một mẫu trong một tập, vì vậy sau khi tìm được một chuỗi trung tâm của mẫu, tất cả các chuỗi tương đồng với chuỗi đó sẽ được loại bỏ. Sau đó, ta tìm mẫu cho tới khi không có chuỗi nào thỏa mãn điều kiện là trung tâm của mẫu.

Trong tập mẫu thu được, có những mẫu chỉ xuất hiện trong số ít trong $\mathbb{A}$. Vì vậy, để đảm bảo tập mẫu bất thường đại diện cho các hành vi đáng ngờ của đối tượng gian lận, ta tiếp tục chọn lọc mẫu dựa trên điểm số của mẫu. Điểm số của một mẫu là tỉ lệ giữa số chuỗi thời gian có chuỗi con tương đồng với chuỗi trung tâm mẫu và tất cả các chuỗi thời gian đang xét. Chỉ có các mẫu có điểm số lớn hơn một ngưỡng nhất định được chọn.

Quy trình khai phá mẫu được trình bày trong thuật toán 1

Thuật toán 1: Thuật toán khai phá mẫu trên chuỗi thời gian.

INPUT: Tập chuỗi thời gian TS, mốc khoảng cách **R**.

OUTPUT Tập mẫu hành vi đáng nghi $\mathbb{P}$

1. Lọc tập chuỗi thời gian của đối tượng gian lận $\mathbb{A}$ từ TS.
2. Mã hoá các chuỗi thời gian trong $\mathbb{A}$ thành chuỗi ký hiệu thu gọn.
3. Xác định các dãy kí tự mã hoá lặp lại trên những chuỗi thời gian trong $\mathbb{A}$ theo từng dạng.
4. Xây dựng ma trận khoảng cách từ các chuỗi con có dãy kí tự mã hoá tìm được ở bước 3.
5. Tìm kiếm chuỗi trung tâm mẫu và các phần tử của mẫu.
6. Loại bỏ các chuỗi con thuộc mẫu vừa tìm được và lặp lại bước 3, nếu không tìm được dãy kí tự lặp lại, chuyển sang bước 7.
7. Dừng thuật toán, kết luận tập mẫu $\mathbb{P}$

Thuật toán 1 trích rút được các thông tin hành vi xuất hiện lặp lại và phổ biến của các đối tượng xấu được gán nhãn bất thường từ trước đó.

Sau khi thực hiện quá trình khai phá mẫu, ta sẽ có các mẫu đại diện cho các hành vi lặp lại trong nhóm các đối tượng gian lận. Ký hiệu tập các mẫu là $\mathbb{P}=\{p_1, p_2, \ldots, p_k\}$. Tập mẫu này sẽ được sử dụng làm cơ sở cho việc phân loại ở bước sau.

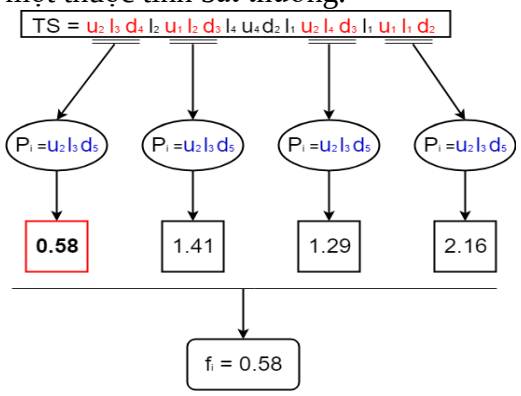
2.4. Phân loại

2.4.1. Thuộc tính bất thường

Xét đối tượng x_i bất kỳ, nếu chuỗi thời gian TS_i tương ứng của đối tượng này chứa nhiều những chuỗi con tương tự với các chuỗi trong tập mẫu bất thường $\mathbb{P}$ thì khả năng x_i từng gian lận là cao. Ngược lại, mức độ nghi ngờ của x_i sẽ thấp.

Với mỗi mẫu p_j thuộc tập mẫu $\mathbb{P}$, độ tương tự giữa p_j với các hành vi thể hiện trên s_i, kí hiệu là $f^i{}_j$, sẽ là khoảng cách tối thiểu giữa p_j đến các chuỗi con của s_i.

Dựa trên khoảng cách từ định nghĩa 4, ta tính khoảng cách p_j với các chuỗi con của TS_i chứa các mẫu giống như p_1 và lấy giá trị nhỏ nhất cho giá trị $f^i{}_1$. Tương tự cho $f^i{}_2, f^i{}_3, \ldots$, ta sẽ có một tập các thuộc tính mới của đối tượng x_i thể hiện sự tương đồng giữa hành vi của đối tượng với hành vi của các đối tượng gian lận. Ký hiệu $F_{new}{}^i = \{f^i{}_1, f^i{}_2, \ldots, f^i{}_k\}$. Hình 5 mô tả một ví dụ tính giá trị của một thuộc tính bất thường.



Hình 5: Tính toán giá trị thuộc tính.

Các thuộc tính của $F_{new}{}^i$ là các số thực không âm biểu diễn độ tương đồng giữa hành vi của đối tượng x_i với hành vi gian lận. Sau đó các thuộc tính này sẽ được sử dụng để phân loại.

2.4.2. Phân loại

Cây quyết định là một trong những thuật toán máy học phổ biến nhất hiện nay và thường được dùng trong bài toán phân lớp với bài toán hồi quy. Thuật toán

này thuộc nhóm các thuật toán học có giám sát.

Cây quyết định là một cấu trúc cây mà mỗi nút biểu diễn một đặc trưng(tính chất), mỗi nhánh biểu diễn một quy luật và mỗi lá biểu diễn một kết quả.

Rừng ngẫu nhiên (Random Forest) là thuật toán phân lớp với nhiều cây quyết định. Mỗi cây quyết định trong rừng được huấn luyện trên các tập con khác nhau của tập luyện sinh ra từ dữ liệu gốc có nhãn bởi phương pháp đóng bao (bagging).

Cho tập luyện $D = \{x_1, x_2, \ldots, x_n\}$. Với $m = 1,2,\ldots,m$, ta chọn tập con ngẫu nhiên ký hiệu $\mathbb{D}_m$và luyện tập $\mathbb{D}_m$ với các cây quyết định DT_m. Sau khi luyện, kết quả dự đoán cho mẫu không nhãn x' được quyết định bởi lấy trung bình dự đoán từ tất cả các cây hồi quy riêng lẻ trên x':

$$\widehat{DT} = \frac{1}{M}\sum_{m=1}^{M} DT_m(x') \qquad (3)$$

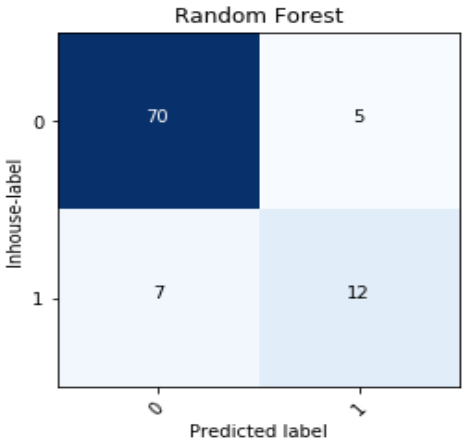
hoặc tổng hợp kết quả của các cây phân loại bằng việc bình chọn theo đa số (majority voting).

2.5. Đánh giá kết quả phân loại bất thường

2.5.1. Ma trận nghi ngờ (Confusion matrix)

Ma trận nghi ngờ cho biết có số lượng phân loại được phân loại đúng và phân loại sai vào từng lớp (bất thường hay không bất thường)

Hình 6 mô tả kết quả phân loại đối tượng vào hai lớp '0' và '1'. Ma trận cho biết rằng có lần lượt 70 đối tượng được phân đúng vào lớp '0',12 đối tượng được phân đúng vào lớp '1' và 12 đối tượng bị phân nhầm lớp (thực tế là lớp '0' nhưng bị phân loại là '1' và ngược lại)



Hình 6: Minh họa ma trận nghi ngờ.

2.5.2. Precision và Recall

Ta xét đến các chỉ số sau:

- TP (True Positive): Số lượng doanh nghiệp được dự đoán là bất thường và thực tế là bất thường.
- FP (False Positive): Số lượng doanh nghiệp được dự đoán là bất thường nhưng thực tế không bất thường.
- FN (False Negative): Số lượng doanh nghiệp được dự đoán là không bất thường nhưng thực tế là bất thường.
- TN (True Negative): Số lượng doanh nghiệp được dự đoán là không bất thường và thực tế đúng là không bất thường.

Khi đó, yếu tố đánh giá Precision được đánh giá bởi công thức:

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

Yếu tố đánh giá Recall được tính bởi công thức:

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

Từ công thức trên, ta có thể thấy yếu tố Precision thể hiện tỷ lệ số lượng dự đoán đúng thực sự là bất thường trên số lượng dự đoán bất thường bởi mô hình.

Yếu tố Recall thể hiện tỷ lệ số lượng dự đoán đúng thực sự là bất thường trên số lượng nhãn bất thường thực tế.

2.5.3. Độ đo F_1

Độ đo F_1 được xác định bởi công thức:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F_1 có giá trị nằm trong $(0,1]$. F_1 càng cao, bộ phân lớp càng tốt.

3. Thí nghiệm và kết quả

3.1. Cài đặt thí nghiệm

3.1.1. Dữ liệu

Dữ liệu được thu thập từ hoạt động mua hàng của khách hàng trong vòng 3 năm. Bộ dữ liệu gồm thông tin giao dịch của 801 khách hàng trong với 70 khách hàng được xác định là có hành vi gian lận trong hoạt động mua hàng, chiếm 8.74%

3.1.2. Mô hình hóa

Dữ liệu được tiền xử lý để thu được các chuỗi thời gian thay đổi địa điểm mua hàng tương ứng với mỗi khách hàng. Từ các khách hàng bị gán nhãn bất thường, ta xây dựng mô hình tìm kiếm mẫu bất thường.

Với giai đoạn $n = 3$ năm, ta phân chia thành 35 phân đoạn bằng nhau và tập $T = \{t_0, t_1, \ldots, t_{35}\}$.

Ta xem xét một doanh nghiệp x_i trong tập các doanh nghiệp $\mathbb{O}$, trong phân đoạn thời gian giữa hai điểm t_j và t_{j+1}, ${G_j}^i$ là tập các địa điểm doanh nghiệp kê khai. Ta có :

$$F : \mathbb{O} \times \Gamma \to \mathbb{R}$$

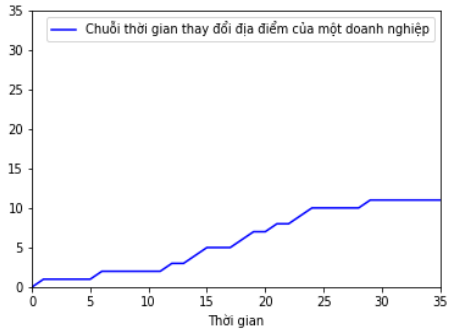
$$\begin{cases} F(x_i, t_{j+1}) = 1 & G^i_{j+1} \not\subseteq G^i_j \\ F(x_i, t_{j+1}) = 0 & otherwise \end{cases}$$

Hàm F đại diện số thay đổi địa điểm kê khai hoạt động xuất nhập khẩu trong giai đoạn $(t_j, t_{j+1}]$.

TS_i là chuỗi thời gian của doanh nghiệp x_i có:

$$TS_i = \{TS^i_0, TS^i_1, \ldots, TS^i_{35}\}$$

$$TS^i_j = \sum_{l=0}^{j} F(x_i, t_l)$$



Hình 7: Chuỗi thời gian thể hiện sự thay đổi địa điểm của một doanh nghiệp trong 3 năm.

Sau khi xây dựng tập chuỗi thời gian cho từng doanh nghiệp, sử dụng thuật toán 1, ta thực hiện quá trình khai phá mẫu. Từ thực nghiệm, ngưỡng điểm số mẫu là 20% cho kết quả tối ưu (hay ta chỉ xét các mẫu xuất hiện trong ít nhất 20% chuỗi thời gian trong $\mathbb{A}$).

Áp dụng với bộ dữ liệu hoạt động mua hàng, nếu ta chọn ngưỡng tương đồng là 0.75, sẽ thu được 30 mẫu bất thường. Một vài mẫu được đưa ra như bảng 2.

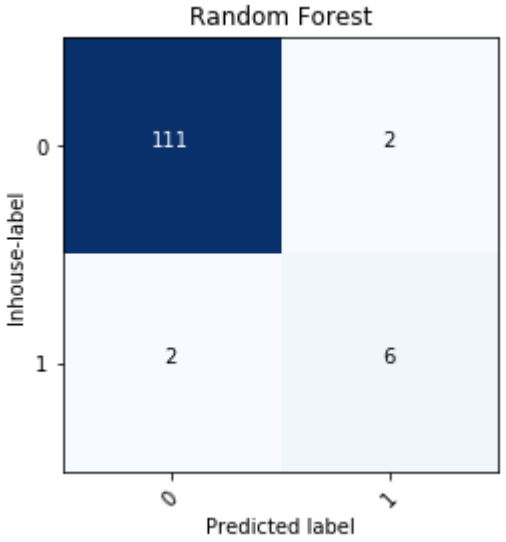
Dạng	Mẫu(SSS)	Điểm số của mẫu(%)
u	u_2	81.82

u	u_3	45.46
u	u_4	25.45
ul	$u_1 l_7$	75.81
ul	$u_2 l_2$	72.73
ul	$u_2 l_5$	38.18
ul	$u_4 l_1$	21.82
ul	$u_3 l_3$	16.36
ul	$u_2 l_8$	18.18
lu	$l_2 u_2$	60
lu	$l_5 u_2$	29.09
lu	$l_8 u_2$	21.82
lu	$l_1 u_4$	23.64
lu	$l_3 u_3$	21.82

Bảng 2: Dãy một số mẫu các hành vi đáng nghi

3.2. Kết quả

Ta tính toán các giá trị của thuộc tính mới lập một tập giá trị thuộc tính tương ứng với mỗi doanh nghiệp. Sau đó sử dụng thuật toán rừng ngẫu nhiên để huấn luyện dữ liệu. Hình 8 mô tả kết quả phân lớp theo ma trận nghi ngờ



Hình 9: Ma trận nghi ngờ của tập kiểm thử gồm 121 doanh nghiệp.

Precision	**Recall**	**F_1-score**
0.75	0.75	0.75

Bảng 3: Bảng độ đo F_1, precision và recall

Trong bảng 3, mặc dù tỷ lệ khách hàng gian lận là nhỏ (chiếm 6.6 %) nhưng mô hình phát hiện được 75% khách hàng gian lận.

4. Kết luận

Trong bài báo này, tôi đã tìm hiểu và nghiên cứu về bài toán phát hiện bất thường, khai phá mẫu bất thường trong chuỗi thời gian, xây dựng mô hình nhận dạng mẫu bất thường và ứng dụng học máy vào việc phát hiện khách hàng bất thường trong dữ liệu khách hàng bán lẻ. Kết quả đạt được của bài báo này như sau:

- Đề xuất kỹ thuật khai phá mẫu bất thường.
- Xây dựng mô hình khai phá mẫu và sử dụng thuật toán rừng ngẫu nhiên để phân lớp khách hàng từ đó tìm ra những khách hàng có hành vi mua hàng bất thường.

5. Lời cảm ơn

Tôi xin có lời cảm ơn tới Viện nghiên cứu Ứng dụng Công nghệ CMC vì những đóng góp và hỗ trợ trong bài báo.

Tài liệu tham khảo

[1] E. Ngai, Y. Hu, Y. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems 50 (3) (2011) 559–569, on quantitative methods for detection of financial fraud. doi:https://doi.org/10.1016/j.dss.2010.08.006.
URL https://www.sciencedirect.com/science/article/pii/S0167923610001302

[2] J. Wu, W. Zeng, Z. Chen, X.-F. Tang, Hierarchical temporal memory method for time-series-based anomaly detection, in: 2016 IEEE 16th International Conference on Data Mining Work-shops (ICDMW), 2016, pp. 1167–1172. doi:10.1109/ICDMW.2016.0168.

[3] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, O. Caelen, Sequence classification for credit-card fraud detection, Expert Systems with Applications 100 (2018) 234–245. doi: https://doi.org/10.1016/j.eswa.2018.01.037.
URL https://www.sciencedirect.com/science/article/pii/S0957417418300435

[4] P. Rousseeuw, D. Perrotta, M. Riani, M. Hubert, Robust monitoring of time series with application to fraud detection, Econometrics and Statistics 9 (2019) 108–121.doi:https://doi.org/10.1016/j.ecosta.2018.05.001.
URL https://www.sciencedirect.com/science/article/pii/S2452306218300303

[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, Data mining for credit card fraud: A comparative study, Decision Support Systems 50 (3) (2011) 602–613, on quantitative methods for detection of financial fraud. doi: https://doi.org/10.1016/j.dss.2010.08.008.
URL https://www.sciencedirect.com/science/article/pii/S0167923610001326

[6] B. Chiu, E. Keogh, S. Lonardi, Probabilistic discovery of time series motifs, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, Association for Computing Machinery, New York, NY, USA, 2003, p. 493–498. doi: 10.1145/956750.956808.
URL https://doi.org/10.1145/956750.956808

[7] Y. Tanaka, K. Iwamoto, K. Uehara, Discovery of time-series motif from multi-dimensional data based on mdl principle, Machine Learning 58 (2005) 269–300.

[8] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140. doi:10.1023/A:1018054314350.
URL https://doi.org/10.1023/A:1018054314350

[9] L. Breiman, Random forests, Machine Learning 45 (2004) 5–32.

Giới thiệu tác giả:

Tác giả Phạm Ngọc Quang Anh tốt nghiệp đại học ngành Toán tin Trường Đại học Bách khoa Hà Nội năm 2020; nhận bằng Thạc sĩ năm 2022 ngành Toán tin. Tác giả là nghiên cứu viên của Viện ứng dụng công nghệ CMC.

Lĩnh vực nghiên cứu: Phát hiện bất thường, hệ khuyến nghị.

Tác giả Vũ Thành Nam là tiến sĩ Toán tin Trường Đại học Bách Khoa Hà Nội. Tác giả là trưởng bộ môn Toán –Tin, Trường Đại học Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Mã hóa bảo mật, Blockchain, lý thuyết mã.

Tác giả Hoàng Văn Đông tốt nghiệp đại học ngành Toán tin Trường Đại học Bách khoa Hà Nội năm 2018; nhận bằng Thạc sĩ năm 2019 ngành Toán tin. NCS tại đại học Thâm Quyến, Trung Quốc. Tác giả đồng thời là giảng viên khoa CNTT, Đại học Thuỷ Lợi.

Lĩnh vực nghiên cứu: AI trong phát hiện bất thường, Phương pháp số, cấu trúc dữ liệu và giải thuật.

Tác giả Lê Anh Ngọc tốt nghiệp đại học ngành toán và tin học tại Trường Đại học Vinh và Trường Đại học Khoa học tự nhiên – Đại học Quốc gia Hà Nội các năm 1996 và 1998. Nhận bằng Thạc sĩ Công nghệ thông tin tại Trường Đại học Bách Khoa Hà Nội năm 2001; nhận bằng Tiến sĩ tại Đại học Quốc gia Kyungpook – Hàn Quốc, chuyên ngành kỹ thuật thông tin và truyền thông năm 2009. Hiện nay tác giả là giảng viên và Giám đốc Swinburne Innovation Space tại Swinburne Việt Nam thuộc Đại học FPT.

Hướng nghiên cứu chính: Hệ thống thời gian thực, mạng truyền thông, Internet of Things, các hệ thống thông minh và IoT.

Tác giả Nguyễn Thị Ngọc Anh, giảng viên Bộ môn Toán ứng dụng, Viện Toán ứng dụng và Tin học, Trường Đại học Bách khoa Hà Nội (SAMI-HUST). Tác giả nhận bằng Tiến sĩ Khoa học Máy tính tại Đại học Pierre et Marie Curie (Paris 6), Cộng hòa Pháp năm 2014.

Lĩnh vực phân tích dữ liệu, mô hình và mô phỏng, dự báo, phát hiện bất thường.