ISSN: 1859-2171
e-ISSN: 2615-9562
TNU Journal of Science and Technology 225(06): 157 - 164

ỨNG DỤNG PHÂN TÍCH DỮ LIỆU VÀ PHÂN LỚP GIÁM SÁT NAIVE BAYES PHÁT HIỆN GIAN LẬN TRONG THANH TOÁN TRỰC TUYẾN

Mai Mạnh Trừng[1], Lê Trung Thực[2*], Đào Thị Phương Anh[1]
[1]Trường Đại học Kinh tế Kỹ thuật Công nghiệp, [2]Trường Đại học Công nghệ Đông Á

TÓM TẮT

Sự phát triển nhanh chóng của giao dịch thanh toán trực tuyến kéo theo tấn công gian lận trong hình thức giao dịch này tăng theo, gây tổn thất to lớn cho nhiều cá nhân, tập thể trong ngành tài chính. Gian lận giao dịch tín dụng trong thanh toán trực tuyến là một trong những hoạt động phi pháp phổ biến và đáng lo ngại nhất. Việc phát hiện, ngăn chặn các hoạt động gian lận giao dịch thông qua phân tích, khai phá dữ liệu kết hợp sử dụng thuật toán học máy là một trong những phương pháp nổi bật hiện nay. Kỹ thuật khai phá dữ liệu được sử dụng để nghiên cứu các mẫu, đặc điểm, thuộc tính, hành vi của giao dịch bình thường, giao dịch bất thường (giao dịch gian lận) dựa trên dữ liệu chuẩn hóa và dữ liệu bất quy tắc. Thuật toán học máy phân lớp nhằm dự đoán, phát hiện giao dịch bình thường, giao dịch gian lận một cách tự động mỗi khi có giao dịch mới phát sinh. Bài viết này nghiên cứu về một số thuật toán học máy có giám sát: Sử dụng mạng Bayes, cây tăng cường Naïve Bayes (Tree Augmented Naïve Bayes – TAN) và Naïve Bayes trong bài toán phân lớp nhị phân dựa trên dữ liệu là hơn 4 triệu bản ghi giao dịch tín dụng trực tuyến tương ứng với khoảng 80 nghìn mã thẻ nhằm phát hiện giao dịch gian lận. Sau khi tiền xử lý dữ liệu bằng phương pháp chuẩn tắc và phân tích thành phần chính (Principal Component Analysis-PCA), tất cả các thuật toán phân lớp đạt độ chính xác hơn 95% so với bộ dữ liệu chưa qua tiền xử lý.

Từ khóa*: Gian lận giao dịch tín dụng; TAN; PCA; Naive bayes, cây tăng cường; mạng Bayes*

Ngày nhận bài: 11/3/2020; Ngày hoàn thiện: 04/5/2020; Ngày đăng: 11/5/2020

DATA ANALYSIS APPLICATION AND NAÏVE BAYES SUPERVISED CLASSIFICATION IN ONLINE PAYMENT

Mai Manh Trung[1], Le Trung Thuc[2*], Dao Thi Phuong Anh[1]
[1]University of Economics Technology for Industries , [2]East Asia University of Technology

ABSTRACT

The fast development of online payment transactions has led to an increase in fraud in this type of transaction, causing great losses for many individuals and collectives in the financial industry. Credit transaction fraud in online payment is one of the most common and disturbing illegal activities. The detection, prevention of fraudulent transactions through analysis and data mining combined using machine learning algorithms is one of the current prominent methods. Data mining techniques are used to study patterns, characteristics, attributes and behaviors of normal transactions, abnormal transactions (fraudulent transactions) based on standardized and irregular data. Class machine learning algorithm to predict, detect normal transactions, fraudulent transactions automatically whenever a new transaction arises. This paper looks at some supervised machine learning algorithms: Using Bayes network, Tree Augmented Naïve Bayes (TAN) and Naïve Bayes in the binary classification problem based on data are more than 4 million online credit transaction records equivalent to about 80,000 card codes to detect fraudulent transactions. After pre-processing the data using the Principal Component Analysis (PCA) method, all classification algorithms achieve 95% more accuracy than the pre-pretreated data set.

Keywords: *Credit transaction fraud; TAN; PCA; Naive bayes; Reinforced trees; Bayes network*

Received: 11/3/2020; Revised: 04/5/2020; Published: 11/5/2020

* Corresponding author. *Email: thuclt12a@gmail.com*

1. Giới thiệu

Theo báo cáo thanh toán quốc tế hàng năm trên Global Payments Report, thẻ tín dụng là phương thức thanh toán trực tuyến được dùng nhiều nhất trên thế giới trong những năm gần đây so với các phương thức khác như sử dụng ví điện tử hay chuyển khoản qua ngân hàng trực tuyến (Internet Banking). Các dịch vụ giao dịch lớn thường bị tội phạm mạng để mắt đến và thực hiện tấn công nhằm gian lận giao dịch thẻ tín dụng. Gian lận thẻ tín dụng được hiểu là việc sử dụng giao dịch một cách trái phép, hành vi giao dịch có gian lận hoặc giao dịch của mã thẻ không hoạt động. Có 3 loại gian lận thẻ tín dụng phổ biến: Gian lận thông thường (đánh cắp, giả mạo), gian lận trực tuyến (các hành vi giao dịch trực tuyến trái phép) và gian lận liên quan đến việc cấu kết giữa các thương gia [1].

Những năm gần đây, gian lận thẻ tín dụng phát triển đến mức đáng báo động. Theo báo cáo của Nilson, tổn thất gian lận thẻ tín dụng toàn cầu đạt 16,31 tỷ đô trong năm 2014 và ước tính sẽ vượt mức 35 tỷ đô vào năm 2022 [2]. Do đó, việc phát triển kỹ thuật phát hiện và ngăn chặn gian lận thẻ tín dụng là cần thiết để chống lại hoạt động phi pháp tài chính này.

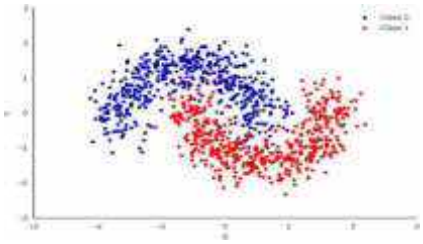
Kỹ thuật phát hiện gian lận thẻ tín dụng được biết đến là quá trình phân lớp, xác định xem một giao dịch tín dụng có phải là gian lận hay không. Phương pháp khai phá dữ liệu kết hợp cùng các thuật toán học máy ngày nay được sử dụng rộng rãi để chống lại các hành vi thám mã trực tuyến nói chung. Trong bài báo, tác giả dùng cách tiếp cận này để phát hiện ra giao dịch tín dụng gian lận. Tác giả ứng dụng khai phá dữ liệu để xác định các mẫu và mô hình từ lượng lớn dữ liệu đã có. Khả năng trích xuất thông tin của khai phá dữ liệu từ tập dữ liệu quy mô lớn sử dụng các kỹ thuật thống kê và toán học sẽ hỗ trợ phát hiện gian lận thẻ tín dụng dựa trên việc phân biệt các đặc điểm của giao dịch bình thường và giao dịch gian lận. Trong khi kỹ thuật khai phá dữ liệu tập trung vào việc tìm ra những thông tin có giá trị, thì thuật toán học máy sẽ tập trung vào việc xây dựng, trích chọn, nghiên cứu các đặc trưng của dữ liệu, từ đó phát triển mô hình nhằm phân lớp, phân cụm dữ liệu.

Ứng dụng của các thuật toán học máy trải rộng trên hầu hết mọi lĩnh vực khoa học máy tính như: Lọc thư rác, tạo chiến dịch quảng cáo online theo thói quen người dùng, chấm điểm tín dụng, phát hiện gian lận giao dịch cổ phiếu, và nhiều ứng dụng khác. Nổi bật trong lĩnh vực học máy này là bài toán phân lớp, bài toán này được giải quyết bằng cách xây dựng, phát triển một mô hình học máy từ mẫu dữ liệu đầu vào, mô hình này sẽ được sử dụng để dự đoán hoặc quyết định cho các dữ liệu đầu vào tiếp theo một cách linh hoạt, tự động thay vì hoạt động như một chương trình lập trình sẵn theo từng trường hợp cụ thể. Có rất nhiều phương pháp học máy khác nhau để xử lý các bài toán khác nhau. Trong bài viết này, chúng tôi tập trung vào thuật toán học máy có giám sát đối với bài toán phân lớp nhị phân, phân lớp mỗi giao dịch tín dụng vào hai lớp, giao dịch bình thường hoặc giao dịch gian lận.

2. Cơ sở lý thuyết

Có khá nhiều các nghiên cứu tận dụng thế mạnh của kỹ thuật khai phá dữ liệu, thuật toán học máy ngăn chặn các hành vi gian lận giao dịch thẻ tín dụng. Ứng dụng kỹ thuật khai phá dữ liệu SOM (Self-Organizing Map) và mạng Nơ-ron [3] cho kết quả lên đến 95% các trường hợp gian lận được dự đoán chính xác. Mô hình Markov ẩn cũng được áp dụng trong phát hiện gian lận thẻ tín dụng với tỷ lệ dự đoán sai giao dịch gian lận khá thấp [4]. Tuy vậy, quá trình chuyển đổi trạng thái khác nhau và việc tính toán xác suất trong mô hình Markov ẩn rất phức tạp và tiêu tốn tài nguyên. Thay vì sử dụng phân lớp dữ liệu, một số nghiên cứu phát hiện gian lận thẻ tín dụng đi theo hướng tiếp cận đó là học phương pháp học dựa trên các thuật toán học máy có giám sát. Nhóm của S.J. Stolfo nghiên cứu hệ thống phát hiện gian lận giao dịch thẻ tín dụng bằng thuật toán cây quyết định ID3, cây

phân lớp hồi quy (CART) [5]. Ý tưởng của hệ thống này là đưa ra giả thiết rằng phân bổ 50/50 giữa trường hợp giao dịch bình thường và giao dịch gian lận, nghiên cứu chỉ ra rằng học phương pháp học sử dụng định lý Bayes làm cơ sở có thể đưa đến kết quả dự đoán đúng giao dịch gian lận rất tốt, nhưng đây không phải là tình huống thực tế, khi mà số lượng giao dịch bình thường có tỷ lệ cao hơn hẳn giao dịch gian lận. Các nhà nghiên cứu khác tiếp cận theo hướng học phương pháp học phân lớp khác như: Sen, Sanjay Kumar, Dash và Sujatha cũng đạt được nhiều kết quả khả quan [6].



Hình 1. *Phân lớp nhị phân*

Bài toán phân lớp (classification) – một trong những bài toán lớn của lĩnh vực học máy được minh họa như hình 1. Nó là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp. Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Có thể hiểu quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu. Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phân lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào. Một số loại học máy được biết đến là học có giám sát, học bán giám sát, học không giám sát, học củng cố hay học phương pháp học. Bài viết này, tác giả tập trung vào học máy có giám sát. Trong các nghiên cứu về bài toán phân lớp, thuật toán học máy có giám sát thường được đánh giá cao vì khả năng kiểm soát các phân lớp thể hiện với sự can thiệp của con người, phân lớp thể hiện sẽ được gán nhãn trước khi đưa vào thuật toán phân lớp. Sau đó, hiệu suất của thuật toán phân lớp sẽ được đánh giá thông qua một số chỉ số nhất định. Cụ thể trong bài toán ngăn chặn tấn công gian lận, tác giả sử dụng phân lớp nhị phân cho dữ liệu vào một trong hai lớp: giao dịch bình thường và giao dịch gian lận [6], [7]. Để xây dựng được mô hình phân lớp và đánh giá được mô hình chúng ta phải trải qua các quá trình như sau:

Bước 1: Chuẩn bị tập dữ liệu huấn luyện và rút trích đặc trưng. Công đoạn này được xem là công đoạn quan trọng trong các bài toán học máy. Nó là đầu vào (input) cho việc học để tìm ra mô hình của bài toán. Chúng ta phải biết cần chọn ra những đặc trưng (thuộc tính) đủ tốt của dữ liệu, lược bỏ những thuộc tính không tốt, gây nhiễu và ước lượng số chiều của dữ liệu bao nhiêu là tốt. Số chiều quá lớn gây khó khăn cho việc tính toán, nhưng cũng không nên giảm thiểu quá mức vì ảnh hưởng đến độ chính xác của dữ liệu.

Bước 2: Xây dựng mô hình phân lớp. Mục đích của mô hình huấn luyện là tìm ra hàm $f(\mathbb{x})$ và thông qua hàm f tìm được nhằm gán nhãn cho dữ liệu. Bước này thường được gọi là học hay huấn luyện:

$$f(\mathbb{x}) = y \quad (1)$$

Trong đó: $\mathbb{x}$ là các véc-tơ đầu vào của dữ liệu, y là nhãn phân lớp hay đầu ra. Thông thường để xây dựng mô hình phân lớp cho bài toán này sử dụng các thuật toán học giám sát như: KNN, mạng nơ-ron, SVM, cây quyết định, Naïve Bayes...

Bước 3: Kiểm tra dữ liệu với mô hình. Sau khi đã tìm được mô hình phân lớp ở bước 2, thì ở bước này công việc là đưa vào các dữ liệu mới để kiểm tra trên mô hình phân lớp.

Bước 4: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất. Quá trình thực hiện bài toán phân lớp qua 4 bước như hình 2. Bước cuối cùng là thực hiện đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu kiểm thử và dữ liệu huấn luyện thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn cần phải thay đổi các tham số của các thuật toán học máy để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. Cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán.

Hình 2. *Quá trình thực hiện bài toán phân lớp*

Mạng Bayes là một trong những kỹ thuật phân lớp được sử dụng rộng rãi nhất trong việc phát hiện gian lận giao dịch thẻ tín dụng trực tuyến. Maes.et.al [8] đã thử nghiệm và đưa ra các chỉ số TP (True Positive), FP (False Positive) của mô hình tạo ra bởi mạng Bayes và mạng Nơ-ron nhân tạo trong bài toán phát hiện gian lận giao dịch thẻ tín dụng. Trong nghiên cứu đó, mạng Bayes cho hiệu suất cao hơn mạng Nơ-ron nhân tạo khoảng 8%, đồng thời có thời gian xử lý ngắn hơn [9]. Thay vì phân tích bằng các phương pháp phân lớp truyền thống, nghiên cứu cuả A.C. Bahnsen đã phát triển một phương pháp phát hiện gian lận thẻ tín dụng dựa trên giá trị rủi ro tối thiểu Bayes (Bayes Minimum Risk) [10]. Ở nghiên cứu này, tác giả xây dựng mô hình phân lớp dựa trên thuật toán phân lớp như: mạng Bayes, cây tăng cường Naïve Bayes (TAN), và Naïve Bayes.

Mạng Bayes là cách biểu diễn đồ thị của sự phụ thuộc thống kê trên một tập hợp các biến ngẫu nhiên, trong đó các nút đại diện cho các biến, còn các cạnh đại diện cho các phụ thuộc có điều kiện. Phân phối xác suất đồng thời của các biến được xác định bởi cấu trúc đồ thị của mạng. Nếu có một cạnh từ nút A tới nút B, thì biến B phụ thuộc trực tiếp vào biến A, và A được gọi là cha của B. Nếu với mỗi biến $x_i, i \in \{1,2,\dots,N\}$ tập hợp các biến cha được ký hiệu bởi $P(x_i)$, thì phân phối có điều kiện phụ thuộc của các biến là tích của các phân phối địa phương:

$$\Pr(x_1, x_2, \dots, x_N) = \prod_{i=1}^{N} \Pr(x_i \mid P(x_i)) \quad (2)$$

Nếu x_i không có cha, ta nói rằng phân phối xác suất địa phương của nó là không có điều kiện, ngược lại thì gọi là có điều kiện. Mạng Bayes có một số lợi thế như khả năng xử lý các đầu vào không hoàn chỉnh, việc học về mối quan hệ nhân quả [11]. Xét bài toán classification với C lớp 1, 2, 3, … , C. Giả sử có một điểm dữ liệu x ∈ R^d. Tính xác suất để điểm dữ liệu này rơi vào phân lớp c, nói cách khác là việc thực hiện tính: $p(y = c|x)$. Hoặc viết gọn thành $p(c|x)$. Đồng nghĩa với tính xác suất để đầu ra là phân lớp c biết rằng đầu vào là một véc-tơ x. Biểu thức này, nếu tính được, sẽ xác định được xác suất để điểm dữ liệu rơi vào mỗi phân lớp. Từ đó có thể xác định phân lớp của điểm dữ liệu đó thuộc vào bằng cách chọn ra phân lớp có xác suất cao nhất:

$$c = \arg \max_{c \in \{1,2,3,\dots c\}} p(c \mid x) \quad (3)$$

Biểu thức này rất khó để tính trực tiếp, áp dụng định lý Bayers:

$$c = \arg \max_c p(c \mid x) = \arg \max_c \frac{p(x|c)p(c)}{P(x)} \quad (4)$$

Do mẫu số p(x) không phụ thuộc vào c nên ta có:

$$c = \arg \max_c p(x|c)p(c) \quad (5)$$

$p(c)$ được hiểu là xác suất một điểm dữ liệu rơi vào phân lớp c. Giá trị này có thể tính bằng MLE (Maximum Likelihood Estimation), tức tỷ lệ số điểm dữ liệu trong tập huấn luyện rơi vào phân lớp c này chia cho tổng số lượng dữ liệu của tập huấn luyện, hoặc cũng có thể đánh giá bằng ước lượng MAP (Maximum a Posteriori). Thành phần còn lại $p(x|c)$, là phân phối của các điểm dữ liệu thuộc vào phân lớp c, để tính toán giá trị này là không dễ dàng do x là biến ngẫu nhiên nhiều chiều, cần rất nhiều dữ liệu huấn luyện mới có thể xây dựng phân phối đó. Giả sử các thành phần của biến ngẫu nhiên x độc lập nhau nếu biết c, khi đó:

$$P(x|c) = p(x_1, x_2, \dots, x_d|c) \prod_{i=1}^{N} P(x_i|c) \quad (6)$$

Giả thiết Naïve Bayes về sự độc lập của số chiều dữ liệu. Với giả thiết này đã tận dụng tối đa tính đơn giản, do đó phân lớp Naïve Bayes có tốc độ huấn luyện và kiểm thử mô hình rất nhanh. Tại bước huấn luyện, các phân phối $p(c)$ và $p(x_i|c)$, i = 1, 2,… , d được xác định dựa vào việc huấn luyện dữ liệu, sử dụng MLE hoặc MAP để tính toán. Tiếp theo, tại bước kiểm thử mô hình với tập dữ liệu

kiểm thử, với mỗi điểm dữ liệu mới x, phân lớp của nó sẽ được xác định bởi:

$$c = \arg \max_{c\in\{1,2,3,\dots c\}} p(c) \prod_{i=1}^{d} p(x_i|c) \quad (7)$$

Việc tính toán $p(x_i|c)$ phụ thuộc hoàn toàn vào loại dữ liệu đầu vào, có ba mô hình Bayes thường được sử dụng bao gồm:

Mô hình Gau-xơ Naïve Bayes. Mô hình này được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục. Với mỗi chiều dữ liệu i và một phân lớp c, x_i tuân theo một phân phối chuẩn có kỳ vọng μ_{ci} và phương sai σ^2_{ci}

Mô hình Naïve Bayes đa thức. Mô hình này chủ yếu được sử dụng trong phân lớp văn bản mà véc-tơ đặc trưng được tính bằng BOW (Bags of Words). Lúc này, mỗi văn bản được biểu diễn bởi một véc- tơ có độ dài d chính là số từ trong từ điển. Giá trị của thành phần thứ i trong mỗi véc-tơ chính là số lần từ thứ i xuất hiện trong văn bản đó. Khi đó, $p(x_i|c)$ tỷ lệ với tần suất từ thứ i (hay đặc trưng thứ i cho trường hợp tổng quát) xuất hiện trong các văn bản của phân lớp c. Giá trị này có thể được tính bằng cách:

$$\lambda_{ci} = p(x_i|c) = \frac{N_{ci}}{N_c} \quad (8)$$

Trong đó:

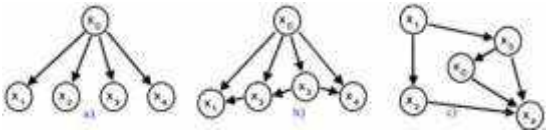
- N_{ci} là tổng số lần từ thứ i xuất hiện trong các văn bản của phân lớp c, nó được tính là tổng của tất cả các thành phần thứ i của các véc-tơ đặc trưng ứng với phân lớp c.

- N_c là tổng số từ (kể cả lặp) xuất hiện trong phân lớp c. Nói cách khác, nó bằng tổng độ dài của toàn bộ các văn bản thuộc vào phân lớp c.

Mô hình Bernoulli Naïve Bayes. Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị nhị phân – bằng 0 hoặc 1. Ví dụ: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không khi đó, $p(x_i|c)$ được tính bằng:

$$p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c)^{1-x_i} \quad (9)$$

Với p(i|c) có thể được hiểu là xác suất từ thứ i xuất hiện trong các văn bản của phân lớp c.



Hình 3. *Mô phỏng cấu trúc của Naïve Bayes (a), TAN(b) và mạng Bayes (c)*

Như hình 3, có sự khác biệt nhỏ giữa Naïve Bayes, TAN và mạng Bayes. Naïve Bayes là một thuật toán phân lớp rất phổ biến vì nó đơn giản, hiệu quả và mang lại hiệu suất tốt trong việc giải quyết các bài toán thực tiễn. Mặt khác, TAN sử dụng hàm tính điểm của Bayes để phát triển mạng Bayes. TAN cho phép tạo ra các cung giữa các nút con x_C (hình 3). Do đó, trình phân lớp TAN có thể tính xác suất từ mỗi nút con và cuối cùng xác định các phân lớp thích hợp với nút con dựa trên xác suất tính toán đó. Mặc dù thông tin được truyền tải bởi TAN có vẻ tốt hơn Naïve Bayes, nhưng hiện chưa có nghiên cứu nào từng thử nghiệm hiệu suất của TAN đối với việc phát hiện gian lận giao dịch thẻ tín dụng.

3. Phương pháp và công cụ

3.1. Giả thuyết

Tham khảo từ các nghiên cứu trước đây, hai kết luận chính được đưa ra để đánh giá việc phát hiện gian lận giao dịch thẻ tín dụng: Kết luận đầu tiên, là dữ liệu thẻ tín dụng đóng vai trò thiết yếu trong việc xác định các đặc trưng của giao dịch gian lận và giao dịch bình thường. Tuy nhiên, quá trình lấy dữ liệu liên quan đến gian lận giao dịch thẻ tín dụng thực sự rất khó khăn do tính bảo mật và nhạy cảm của dữ liệu. Do đó, nếu không thể thu thập được bộ dữ liệu thực tế đủ lớn, các nhà nghiên cứu bắt buộc phải xây dựng mô phỏng dữ liệu thực tế. Để làm được điều đó các tác giả của những nghiên cứu này đã sử dụng dữ liệu được tạo ra dựa trên một số đặc điểm được cho là có tác động đáng kể đến việc phát hiện gian lận. Ví dụ: Nếu khách hàng nhập sai mã pin nhiều lần hoặc địa chỉ giao hàng thực

tế khác với địa chỉ thanh toán hoặc ngày và thời gian giao dịch quá sát nhau trong khi số lượng giao dịch lại lớn hơn hẳn so với những hoạt động trước đó, thì đó có thể được quy thành giao dịch khả nghi. Vì vậy, dữ liệu mô phỏng được phát triển với một số thuộc tính như: Số thẻ tín dụng, số tham chiếu giao dịch, mã thiết bị thực hiện giao dịch, mã pin thực tế, mã pin đã nhập, lượng tiền giao dịch, ngày giao dịch, thời gian, địa điểm giao dịch, địa chỉ thanh toán và địa chỉ giao hàng… Kết luận thứ hai, là hầu hết các nghiên cứu trước đây đã cố gắng sử dụng các phân lớp bất đồng bộ để đo lường hiệu suất phát hiện giao dịch gian lận hay giao dịch bình thường. Với ý định đóng góp thêm cho nền tảng kiến thức, thí nghiệm thứ hai được thực hiện để đánh giá hiệu suất của các phân lớp được đưa ra trong việc phân lớp các hoạt động gian lận thẻ tín dụng. Do đó, các giả thuyết thứ nhất và thứ hai phản ánh hai thí nghiệm được nêu như sau:

- Giả thuyết (1): Tập dữ liệu mô phỏng được tạo ra dựa trên các hành vi đáng ngờ có thể được sử dụng để phân lớp trong khai phá dữ liệu.

- Giả thuyết (2): Hiệu suất trên bộ dữ liệu thông qua quá trình tiền xử lý tốt hơn so với tập dữ liệu thô.

3.2. Phương pháp, công cụ

Tổng quan về quy trình thực hiện xây dựng và đánh giá mô hình trong bài báo được minh họa trong hình 4.

Hình 4. *Quy trình xây dựng và đánh giá mô hình*

Chuyển đổi, chuẩn hóa dữ liệu (data transformation) và điều chỉnh giảm dữ liệu (data reduction) là quá trình tiền xử lý dữ liệu. Dữ liệu thô sẽ được làm "sạch" và chuyển đổi thành dạng thích hợp để đánh giá và đưa vào các thuật toán phân lớp. Bước chuẩn hóa, chuyển đổi dữ liệu bao gồm các hoạt động: chuẩn hóa, làm mịn, tổng hợp, xây dựng, trích chọn thuộc tính và khái quát hóa dữ liệu như hình 4. Trong khi đó, bước điều chỉnh giảm dữ liệu lại nhằm vào việc giảm số lượng các thuộc tính bằng cách gộp các thuộc tính đơn lẻ lại với nhau thành thuộc tính tổng hợp, loại bỏ các thuộc tính không liên quan và phân tích thành phần chính. Mục tiêu của việc áp dụng phương pháp này là xác định và giảm tính đa chiều của tập dữ liệu (giảm tính phức tạp tính toán), tận dụng được nhiều hơn ý nghĩa của thuộc tính cơ bản khi chúng kết hợp với nhau. Một trong những ưu điểm của kỹ thuật này đó là trong quá trình giảm tính đa chiều của dữ liệu nhưng không gây ra mất mát đáng kể nào đối với thông tin của dữ liệu.

Hình 5. *Phần mềm nguồn mở WEKA*

Tiếp theo, tác giả sử dụng WEKA (Waikato Environment for Knowledge Analysis) để đo lường hiệu suất của các thuật toán phân lớp. WEKA là một phần mềm học máy được Đại học Waikato, New Zealand phát triển bằng Java giao diện như hình 5. Nó là một công cụ mã nguồn mở nổi bật được sử dụng rộng rãi để nghiên cứu nhiều bài toán thực tế như: Phân tích ý kiến, phát hiện tính cách, loại bỏ thư rác và phát hiện gian lận. Việc phân lớp được thực hiện bằng kỹ thuật xác thực chéo 10 lần. Kỹ thuật này được áp dụng rộng rãi

trong khai phá dữ liệu và học máy do quá trình huấn luyện và kiểm thử được thực hiện trên toàn bộ tập dữ liệu. Bộ dữ liệu được chia thành mười phần, mỗi phần được đưa ra theo lượt và cuối cùng kết quả trung bình được tính toán. Nói cách khác, mỗi điểm dữ liệu trong bộ dữ liệu đã được sử dụng một lần để kiểm thử và 9 lần cho huấn luyện. Sau đó, để đo lường hiệu suất của các thuật toán phân lớp tác giả sử dụng các giá trị sau:

- TP (True Positive) là số lượng giao dịch gian lận được xác định là gian lận.
- FP (False Positive) là số lượng giao dịch bình thường nhưng được xác định là gian lận.
- TN (True Negative) là số lượng giao dịch gian lận được xác định là bình thường.
- FN (False Negative) là số lượng giao dịch bình thường nhưng được xác định là gian lận.

Nghiên cứu này tác giả đánh giá hiệu suất thuật toán phân lớp dựa trên các chỉ số:

- Tỷ lệ chính xác của giao dịch gian lận (TPR – True Positive Rate)
- Tỷ lệ sai lệch của giao dịch gian lận (FPR – False Positive Rate)
- Tỷ lệ dự đoán chính xác (P – Precision)
- Độ tin cậy (A – Accuracy)
- Tốc độ xử lý phân lớp (PS – Processing Speed)

4. Đánh giá kết quả

Nghiên cứu này sử dụng 2 bộ dữ liệu phục vụ 2 trường hợp thử nghiệm. Một là với bộ dữ liệu thô và một là với bộ dữ liệu mới được tạo bằng cách chuyển đổi, chuẩn hóa dữ liệu và điều chỉnh giảm dữ liệu (thông qua tiền xử lý dữ liệu).

4.1. Kết quả thử nghiệm 1

Bảng 1. *Bảng kết quả trường hợp 1*

Tham số	Mạng Bayes	Naïve Bayes	TAN
TPR (%)	33,0	52,3	75,9
FPR (%)	67,0	47,7	24,1
P (%)	22,0	46,0	73,3
PS (giây)	10,08	10,06	55,0
A (%)	43,6	54,0	84,8

Trong thí nghiệm 1, tác giả sử dụng dữ liệu thô với hơn 4 triệu bản ghi giao dịch của khoảng 80 nghìn mã thẻ giao dịch từ một tổ chức tài chính để đánh giá hiệu suất của các mô hình. Kết quả (bảng 1) cho thấy, các chỉ số TPR (75,9%), tỷ lệ dự đoán chính xác P (73,3%) và độ tin cậy A (84,8%) của TAN là cao nhất trong các thuật toán phân lớp. Chỉ số FPR thấp nhất của TAN cho thấy khả năng xử lý dữ liệu thô vượt qua các phân lớp khác, nhưng tốc độ xử lý của nó là 55 giây, chậm hơn so với mạng Bayes (10,08 giây), Naïve Bayes (10,06 giây). Nguyên nhân do quá trình tính xác suất và tạo mô hình cây tăng cường là phức tạp hơn, do đó quá trình xử lý dữ liệu lâu hơn. Để tăng khả năng phân lớp, trong trường hợp thử nghiệm 2, dữ liệu thô sẽ được tiền xử lý bằng các kỹ thuật phân tích, khai phá dữ liệu.

4.2. Kết quả và phân tích thử nghiệm 2

Bảng 2. *Bảng kết quả trường hợp 2*

Tham số	Mạng Bayes	Naïve Bayes	TAN
TPR (%)	90,8	99,4	99,8
FPR (%)	9,2	0,6	0,2
P (%)	92,8	95,0	98,3
PS (giây)	2,01	2,03	31,2
A (%)	95,7	96,9	99,6

Đối với thử nghiệm này, dữ liệu đã được tiền xử lý bằng phương pháp chuẩn hóa và phân tích thành phần chính. Sau khi tiền xử lý dữ liệu, tất cả các thuật toán phân lớp cho kết quả tốt hơn rất nhiều so với bộ dữ liệu thô ban đầu. Kết quả như bảng 2 cho thấy: Tốc độ xử lý nhanh hơn, độ tin cậy cao hơn và chỉ số FPR thấp hơn. Khả năng phân lớp của mạng Bayes cũng cải thiện đáng kể. TPR của các thuật toán tăng gần 200% sau tiền xử lý dữ liệu. Ngoài ra, tốc độ xử lý dữ liệu cũng tăng đáng kể so với bộ dữ liệu thô ở trường hợp 1, và TAN vẫn cho hiệu suất tốt nhất với chỉ số TPR lên đến 99,8%, độ tin cậy là 99,6%, tốc độ xử lý cũng chỉ còn 31,2 giây.

5. Kết luận

Bài báo đã trình bày cơ sở lý thuyết về phân tích dữ liệu và phân lớp giám sát NAÏVE

BAYES. Hai bộ dữ liệu một bộ dữ liệu thô, một bộ dữ liệu mới đã được sử dụng trong thử nghiệm. Kết quả trên bộ dữ liệu mới được chuẩn hóa với các tham số tương ứng tốt hơn nhiều so với bộ dữ liệu thô ban đầu.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1]. N. Sivakumar, and Dr. R. Balasubramanian, "Fraud Detection in Credit Card Transactions: Classification, Risks and Prevention Techniques," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1379-1386, 2015.

[2]. The Nilson Report, "Global Card Fraud Losses Reach $16.31 Billion — Will Exceed $35 Billion in 2020 According to The Nilson Report", August, 2015. [Online]. Available: https://www.businesswire.com/news/home/20150804007054/en/Global-Card-Fraud-Losses-Reach-16.31-Billion. [Accessed Dec. 2019].

[3]. N. Ogwueleka, "Data mining application in credit card fraud detection system," *Journal of Engineering Science and Technology*, vol. 6, no. 3, p. 311, 2011.

[4]. V. Bhusari, and S. Patil, "Application of hidden markov model in credit card fraud detection," *International Journal of Distributed and Parallel Systems (IJDPS)*, vol. 2, no. 6, pp. 203-211, November, 2011.

[5]. S. J. Stolfo, D. W. Fan, W. Lee, A. L. Prodromidis, and P. K. Chan, "Credit card fraud detection using meta-learning: issues and initial results," Proc. AAAI Workshop AI Methods in Fraud, 1998, pp. 83-90.

[6]. S. Y. Sait, M. S. Kumar, and H. A. Murthy, "User traffic classification for proxy-server based internet access control," IEEE 6th International Conference on Signal Processing and Communication Systems (ICSPCS), 2012, pp. 1-9.

[7]. E. M. Carneiro, L. A. V. Dias, A. M. Da Cunha, and L. F. S. Mialaret, "Cluster analysis and artificial neural networks: A case study in credit card fraud detection," 12th ed. International Conference on Information Technology-New Generations, 2015, 122-126.

[8]. S. Maes, K. Tuyls, B. Vanschoenwinkel and B. Manderick, "Credit Card Fraud Detection Using Bayesian and Neural Networks. in Proceedings of the First International NAISO Congress on NEURO FUZZY THECHNOLOGIES," Proceedings of the First International NAISO Congress on NEURO FUZZY THECHNOLOGIES (Havana, Cuba), 2002, pp. 16-19.

[9]. R. Najafi and A. Mohsen, "Network intrusion detection using tree augmented naive-bayes", The Third International Conference on Contemporary Issues in Computer and Information Sciences (CICI), 2012, pp. 396-402.

[10]. R. Jain, B. Gour, and S. Dubey, "A hybrid approach for credit card fraud detection using rough set and decision tree technique," *International Journal of Computer Applications*, vol. 139, no.10, pp. 1-6, 2016.

[11]. A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, "Cost sensitive credit card fraud detection using bayes minimum risk," 12th International Conference on Machine Learning and Applications, 2013.