Journal of Mining and Earth Sciences
Website: http://jmes.humg.edu.vn

Detecting outliers in GNSS position time series using machine learning techniques

Huy Dinh Nguyen, Trong Dinh Tran *

Hanoi University of Civil Engineering, Hanoi, Vietnam

ARTICLE INFO

Article history:
Received 27th Mar. 2023
Revised 30th July 2023
Accepted 24th Aug. 2023

Keywords:
GNSS position time series,
Isolation Fores,
LOF,
O-C SVM,
Outlier.

ABSTRACT

The Global Navigation Satellite System (GNSS) position time series is applied in studies that require high-precision positioning, such as monitoring tectonic movements and Earth deformation. Outliers in GNSS position time series can significantly impact the accuracy of station positioning and movement parameters, leading to distorted data analysis outcomes. This study investigates the effectiveness of three machine learning techniques, including-Isolation Forest, One-Class Support Vector Machines (O-C SVM), and Local Outlier Factor (LOF) for outlier detection in GNSS position time series, with a specific focus on the SYNT model where outliers account for a substantial proportion (15%). Through comprehensive analysis, our results highlight the exceptional performance of the Isolation Forest method. It demonstrates remarkable accuracy in identifying outliers, effectively detecting the majority of them, and achieving an area under the ROC curve close to 1. In contrast, the LOF method performs less effectively in outlier detection, while the O-C SVM method displays relatively higher accuracy in identifying normal data points. These findings emphasize the significant advantages of leveraging machine learning approaches in processing continuous GNSS measurement data. By effectively identifying and handling outliers, these techniques enhance the accuracy and reliability of data analysis in GNSS position time series, ultimately establishing their superiority in the field of data analysis.

*Corresponding author
E - mail: trongtd@huce.edu.vn
DOI: 10.46326/JMES.2023.64(4).03

Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn

Phát hiện ngoại lai trong chuỗi tọa độ GNSS bằng máy học

Nguyễn Đình Huy, Trần Đình Trọng *

Trường Đại học Xây dựng Hà Nội, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 27/3/2023
Sửa xong 30/7/2023
Chấp nhận đăng 24/8/2023

Từ khóa:
Chuỗi tọa độ GNSS,
Isolation Forest,
LOF,
Ngoại lai,
O-C SVM.

TÓM TẮT

Chuỗi tọa độ (position time series) nhận được từ kết quả đo GNSS (Global Navigation Satellite System) liên tục được ứng dụng trong các nghiên cứu yêu cầu định vị độ chính xác cao như dịch chuyển mảng kiến tạo, biến dạng vỏ trái đất,... Ngoại lai hay dị thường (outlier) cần phải loại bỏ trong xử lý số liệu nói chung, đặc biệt trong phân tích chuỗi tọa độ GNSS do chúng làm giảm độ chính xác khi xác định vị trí điểm đo và các tham số dịch chuyển của điểm đo, làm nhiễu kết quả phân tích dữ liệu của chuỗi. Với những ưu điểm vượt trội so với các phương pháp thống kê, hay phương pháp cửa sổ trượt,... trong nghiên cứu này, nhóm nghiên cứu sử dụng 3 phương pháp máy học được đánh giá là tối ưu trong phát hiện ngoại lai là Isolation Forest, One-Class Support Vector Machines (O-C SVM) và Local Outlier Factor (LOF) để phát hiện ngoại lai chiếm tỉ lệ cao (15%) của chuỗi tọa độ GNSS mô hình SYNT. Kết quả cho thấy Isolation Forest đạt hiệu suất tốt nhất, với độ chính xác cao, khả năng tìm ra hầu hết các điểm ngoại lai và diện tích dưới đường cong ROC gần 1, LOF có hiệu suất kém hơn, trong khi O-C SVM chỉ có độ chính xác tương đối cao trong việc xác định các điểm bình thường. Kết quả nghiên cứu góp phần khẳng định ưu điểm vượt trội của các phương pháp máy học trong việc xử lý số liệu đo GNSS liên tục.

1. Mở đầu

Công nghệ định vị GNSS (Global Navigation Satellite System) liên tục cho phép xác định tọa độ các trạm đo liên tục theo theo thời gian, các tọa độ này tạo thành chuỗi theo thời gian (GNSS position time series). Chuỗi tọa độ GNSS được ứng dụng rộng rãi, đặc biệt trong các lĩnh vực yêu cầu định vị độ chính xác cao và liên tục theo thời gian (Bevis và nnk., 2020; Gomo và nnk., 2017; Montillet & Bos, 2020). Trong lĩnh vực nghiên cứu dịch chuyển mảng kiến tạo, biến dạng vỏ trái đất và các hoạt động địa chấn, chuỗi tọa độ GNSS cung cấp vận tốc dịch chuyển (Ito và nnk., 2019; Kall và nnk., 2021; Dinh và nnk., 2023; Tsai và nnk., 2015) và các hiện tượng dịch chuyển bề mặt đất khác (Klos và nnk., 2020; Riel và nnk., 2014), là các tham số quan trọng và chính xác cho các nghiên cứu về động đất, núi lửa các các hiện tượng địa vật lý khác (Métivier và nnk., 2014; Montillet và nnk., 2015; Trọng & Huy, 2023); Trong lĩnh vực nghiên cứu khí hậu, chuỗi tọa độ GNSS liên tục cung cấp các trị đo mực nước biển dâng, sụt lún đất và các dịch chuyển băng hà, cho phép nghiên cứu các tương tác của hệ thống động

*Tác giả liên hệ
E - mail: trongtd@huce.edu.vn
DOI: 10.46326/JMES.2023.64(4).03

của trái đất tới biến đổi khí hậu (Teferle và nnk., 2008);…

Chuỗi tọa độ GNSS của các trạm đo, theo thời gian ngày càng nhiều lên và các giá trị tọa độ ngoại lai hay dị thường (outlier) cũng tăng theo. Các ngoại lai là các giá trị tọa độ bất thường, không tuân theo mô hình thông thường của chuỗi tọa độ (Hawkins, 1980), luôn tồn tại trong chuỗi do nhiều nguyên nhân, như (Tran, 2013; Tran và nnk., 2022): (1) Nhiễu đo: các tác động từ môi trường như nhiễu điện từ, đa đường dẫn, hay động đất có thể gây nhiễu và làm biến đổi trị đo, dẫn đến sự xuất hiện của ngoại lai; (2) Lỗi thiết bị thu GNSS: lỗi hoặc sự không chính xác của các thiết bị thu GNSS có thể gây ra các trị đo bị sai hoặc gián đoạn cũng dẫn đến ngoại lai trong chuỗi tọa độ; (3) Lỗi do thuật toán xử lý: các phương pháp xử lý dữ liệu GNSS có thể gây ra lỗi tính toán hoặc sai sót xử lý dữ liệu, dẫn đến xuất hiện các ngoại lai trong chuỗi tọa độ. Các ngoại lai trong chuỗi tọa độ GNSS dẫn tới: (1) sai lệch vị trí điểm đo xác định được; (2) giảm độ chính xác của chuỗi tọa độ GNSS dẫn đến tính toán các tham số chuyển dịch của điểm đo không chính xác; (3) nhiễu dữ liệu và gây khó khăn cho các quá trình phân tích và xử lý chuỗi tọa độ GNSS. Loại bỏ ngoại lai là một quá trình quan trọng để đảm bảo tính chính xác và tin cậy của dữ liệu GNSS. Việc phát hiện và loại bỏ ngoại lai, có thể kể đến các phương pháp phổ biến như phương pháp thống kê (Hieu và nnk., 2018; Wu và nnk., 2017), phương pháp cửa sổ trượt (Tran và nnk., 2022, 2016), kỹ thuật làm mượt dữ liệu (Trần và nnk., 2014),… Các phương pháp này đạt hiệu quả cao hay không còn tuỳ thuộc vào số lượng, đặc điểm và sự phân bố của các ngoại lai.

Hiện nay, với sự phát triển mạnh mẽ của việc ứng dụng máy học (machine learning), việc ứng dụng trong phát hiện ngoại lai trong chuỗi tọa độ GNSS mang lại nhiều ưu điểm. Từ khả năng học tập và thích ứng, xử lý dữ liệu phức tạp, tính linh hoạt và khả năng mở rộng, hiệu suất cao, đến giảm thiểu công sức và thời gian, các phương pháp máy học đóng góp quan trọng vào việc cải thiện việc phát hiện ngoại lai và nâng cao chất lượng dữ liệu GNSS. Việc ứng dụng máy học trong xử lý chuỗi GNSS nói chung được nhiều học giả trên thế giới quan tâm. Gao và nnk. (2022) đã ứng dụng các phương pháp học máy để xác định mô hình và dự báo chuyển động, trong đó có bước loại bỏ ngoại lai của chuỗi GNSS của 9 trạm GNSS đo liên tục trong 13 năm. Kết quả cho thấy độ chính xác tăng lên rõ rệt so với phương pháp số bình phương nhỏ nhất. Kiani (2020) ứng dụng phương pháp học máy có giám sát trong việc phát hiện ngoại lai và dự báo động đất từ các chuỗi tọa độ GNSS của 845 trạm GEONET, kết quả cho thấy loại bỏ hiệu quả ngoại lai và độ chính xác dự báo động đất đạt cỡ 2h cho động đất tại Tohoku vào năm 2011. Trong khi đó, việc ứng dụng học máy trong xử lý số liệu đo GNSS liên tục ở nước ta còn ít, có thể kể đến nghiên cứu của Phong và nnk. (2023), trong đó các tác giả đã sử dụng mạng nơ-ron nhân tạo để phân tích dự báo chuyển dịch thẳng đứng khu vực đồng bằng sông Cửu Long từ dữ liệu GNSS liên tục của điểm CTHO, kết quả dự báo đạt độ chính xác rất cao. Do các công bố còn ít và kết quả ứng dụng máy học đạt độ chính xác cao như vậy, nên hướng nghiên cứu này là rất triển vọng, nhất là khi mạng lưới VNGEONET, một mạng lưới GNSS liên tục có quy mô lớn và chính xác của nước ta đã được đưa vào sử dụng mới đây.

Trong bài báo này, nhóm nghiên cứu đã ứng dụng ba phương pháp máy học là: phương pháp Isolation Forest, phương pháp One-Class Support Vector Machines (OC-SVM) và phương pháp Local Outlier Factor (LOF), để phát hiện ngoại lai trong chuỗi tọa độ GNSS của một điểm mô hình, từ đó đánh giá hiệu suất, độ chính xác của các phương pháp máy học đó. Các phương pháp này thuộc nhóm các phương pháp máy học không giám sát, được đánh giá là tối ưu trong phát hiện ngoại lai trong tập dữ liệu nói chung.

Trong phần 2 của bài báo sẽ trình bày về các phương pháp máy học này và cách chúng được áp dụng trong việc phát hiện ngoại lai trong chuỗi tọa độ GNSS, đồng thời sẽ tập trung vào thực hiện các thử nghiệm và đánh giá hiệu quả của các phương pháp máy học trong việc phát hiện ngoại lai trong chuỗi tọa độ GNSS.

2. Các phương pháp máy học sử dụng cho nghiên cứu

2.1. Phương pháp Isolation Forest

Đây là phương pháp dựa trên mô hình cây quyết định (decision tree) được sử dụng để phát hiện ngoại lai trong dữ liệu (Liu và nnk., 2008). Phương pháp này giả thiết rằng các điểm ngoại lai có xu hướng bị cô lập hơn so với các điểm bình thường trong không gian dữ liệu, tạo ra các cây quyết định ngẫu nhiên và đo số lần cần thiết để cô

lập một điểm ngoại lai từ dữ liệu. Các điểm ngoại lai cần ít cây để cô lập, trong khi các điểm bình thường cần nhiều cây để cô lập.

Quá trình xây dựng Isolation Forest bao gồm các bước sau:

(1) Chọn một mẫu ngẫu nhiên từ tập dữ liệu và xác định giới hạn trên và dưới của biến trong tập dữ liệu.

(2) Chọn một giá trị ngẫu nhiên trong khoảng giới hạn của biến làm giá trị phân chia (split value).

(3) Chia dữ liệu thành hai phần dựa trên giá trị phân chia.

(4) Lặp lại các bước trên cho đến khi mỗi nút lá chỉ chứa một điểm dữ liệu hoặc đạt được số lượng cây quyết định tối đa.

Sau khi xây dựng cây, phương pháp Isolation Forest tính toán độ cô lập (isolation score) cho mỗi điểm dữ liệu trong dữ liệu dựa trên các cây quyết định. Độ cô lập cho mỗi điểm dữ liệu là số lần phải cắt cây để cô lập điểm đó. Công thức tính toán độ cô lập cho một điểm như sau:

$$s(x,T) = 2^{-\frac{E(h(x))}{c(n)}} \quad (1)$$

Trong đó: $s(x,T)$ - độ cô lập của điểm x trên cây T; $E(h(x))$ - số lần cắt cây để cô lập điểm x trên cây T; $c(n)$ - giá trị trung bình của số lần cắt cây cần thiết để cô lập một điểm trong dữ liệu n điểm.

Độ cô lập của mỗi điểm dữ liệu thường nằm trong khoảng [0,1], trong đó giá trị gần 1 cho biết điểm đó có xu hướng bị cô lập và có khả năng là ngoại lai. Công thức tính toán độ cô lập và quá trình xây dựng cây của Isolation Forest giúp nhanh chóng và hiệu quả trong việc phát hiện điểm ngoại lai trong chuỗi tọa độ GNSS.

2.2. Phương pháp One-Class Support Vector Machines

Phương pháp One-Class Support Vector Machines (OC-SVM) (Schölkopf và nnk., 2001) là một phương pháp học để phát hiện điểm ngoại lai trong dữ liệu. OC-SVM dựa trên giả định rằng các điểm dữ liệu bình thường chiếm phần lớn trong không gian dữ liệu và các điểm ngoại lai là các điểm hiếm gặp hoặc không tuân theo phân phối chung. Phương pháp này xây dựng một ranh giới (boundary) xung quanh các điểm dữ liệu bình thường và điểm dữ liệu nằm bên trong ranh giới được coi là bình thường, trong khi điểm dữ liệu nằm bên ngoài ranh giới được coi là ngoại lai.

Quá trình huấn luyện OC-SVM bao gồm việc tìm ra ranh giới quyết định tốt nhất hay mặt siêu phẳng (hyperplane) để phân cách dữ liệu bình thường và ngoại lai. Hyperplane này được xác định bằng cách tìm nghiệm tối ưu của bài toán tối thiểu hóa hàm mục tiêu.

Hàm mục tiêu của bài toán tối thiểu hóa được xác định bởi công thức sau:

$$\frac{1}{2}w^T.w + v.\frac{1}{2}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i.\phi(x_i) \quad (2)$$

Trong đó: w - vector trọng số của hyperplane; v - tham số điều chỉnh độ lớn của vùng bình thường (nu-parameter); n - số lượng điểm dữ liệu; ξ_i - biến lỏng để cho phép các điểm dữ liệu bị vi phạm (slack variables); α_i - hệ số Lagrange tương ứng với mỗi điểm dữ liệu; $\phi(x_i)$ - hàm kernel tính toán độ tương đồng giữa các điểm dữ liệu x_i.

Sau khi tìm ra nghiệm tối ưu, OC-SVM sử dụng hyperplane và hàm kernel để phân loại các điểm dữ liệu. Các điểm nằm trong vùng bình thường được coi là bình thường, trong khi các điểm nằm bên ngoài vùng bình thường được coi là ngoại lai.

Phương pháp OC-SVM thường được sử dụng trong các bài toán phát hiện ngoại lai, phát hiện bất thường trong dữ liệu, và phân loại một lớp dữ liệu đơn lẻ.

2.3. Phương pháp Local Outlier Factor

Đây là phương pháp phi cơ sở được sử dụng để phát hiện ngoại lai trong dữ liệu - phương pháp Local Outlier Factor (LOF) (Breunig và nnk., 2000). LOF dựa trên nguyên tắc các điểm dữ liệu bình thường thường có mật độ gần nhau, trong khi các điểm ngoại lai thường có mật độ thấp hơn và nằm cách xa các điểm dữ liệu khác. Phương pháp này tính toán hệ số Local Outlier Factor (LOF) cho mỗi điểm dữ liệu, đo lường mức độ "ngoại lai" của điểm đó dựa trên mật độ của các điểm xung quanh:

$$LOF(x) = \frac{\sum_{y\in N_k(x)}\frac{LRD(y)}{LRD(x)}}{\|N_k(x)\|} \quad (3)$$

Trong đó: LRD (Local Reachability Density) - đại lượng thể hiện mật độ của mỗi điểm dữ liệu trong ngữ cảnh của các điểm xung quanh:

$$LRD(x) = (\sum_{y \in N_k(x)} \frac{RD(x,y)}{\|N_k(x)\|})^{-1} \quad (4)$$

Trong đó: $N_k(x)$ - tập các điểm hàng xóm gần nhất của điểm x (k-nearest neighbors); $RD(x,y)$ - khoảng cách tiếp cận (reachability distance) giữa hai điểm x, y (Đây là một giá trị thể hiện mức độ khó khăn trong việc tiếp cận từ điểm x đến y dựa trên khoảng cách của y và mật độ của y trong ngữ cảnh của x).

LOF có giá trị lớn hơn 1 cho các điểm dữ liệu được coi là ngoại lai và giá trị gần bằng 1 cho các điểm dữ liệu bình thường.

Phương pháp LOF giúp xác định các điểm ngoại lai không chỉ dựa trên các thông số thống kê, mà còn dựa trên mối quan hệ giữa các điểm dữ liệu trong không gian. Phương pháp LOF là một công cụ mạnh để phát hiện ngoại lai trong dữ liệu, đặc biệt là khi các điểm ngoại lai không tuân theo phân phối thông thường và không được định nghĩa rõ ràng. Phương pháp này được áp dụng rộng rãi trong các lĩnh vực như phân tích dữ liệu, nhận dạng gian lận, phát hiện xâm nhập và nhiều ứng dụng khác.

3. Số liệu và phương pháp nghiên cứu

3.1. Số liệu

Chuỗi tọa độ GNSS sử dụng cho nghiên cứu là số liệu mô hình của trạm đo SYNT, được tạo thành gồm các tọa độ hàng ngày liên tục trong hơn 3 năm từ 2019.0 đến 2022.5 (1095 tọa độ liên tục). Việc tạo bộ số liệu được thực hiện như sau: (1) tạo chuỗi tọa độ chuyển dịch theo phương trình gồm dịch chuyển tuyến tính và chu kỳ theo mùa (Tran và nnk., 2021); (2) gán sai số ngẫu nhiên theo luật phân phối chuẩn với độ lệch chuẩn ±1 mm vào tất cả các tọa độ của chuỗi; (3) gán sai số thô có biên độ từ ±10 mm đến ±200 mm vào 164 tọa độ trong chuỗi để tạo các ngoại lai (tương đương 15%). Các bước tạo dữ liệu được thực hiện bằng cách lập trình chương trình sử dụng ngôn ngữ Python và các hàm toán học, hàm xác suất thống kê trong thư viện Numpy.

Nhằm đánh giá hiệu quả vượt trội của các phương pháp máy học so với các phương pháp thống kê trong phát hiện ngoại lai, trong nghiên cứu này nhóm nghiên cứu đã gán giá trị lớn, từ ±10 mm đến ±200 mm (thông thường các ngoại lai là các giá trị lớn hơn từ 2÷3 lần độ lệch chuẩn), đồng thời tạo ngoại lai chiếm tỉ lệ 15% là một tỉ lệ lớn mà các phương pháp thống kê không đạt hiệu quả (thông thường các phương pháp thống kê chỉ có hiệu quả khi tỷ lệ ngoại lai nhỏ hơn 5% (Tran, 2013)).

Hình 1 thể hiện chuỗi thành phần tọa độ x (north) của điểm SYNT, trong đó các điểm màu xanh là các điểm tọa độ thông thường, còn các điểm màu đỏ là các tọa độ ngoại lai. Trục tung thể hiện giá trị tọa độ north (North Coordinate) theo đơn vị mm, trục hoành thể hiện thời gian của chuỗi theo đơn vị năm (Time).

3.2. Phương pháp

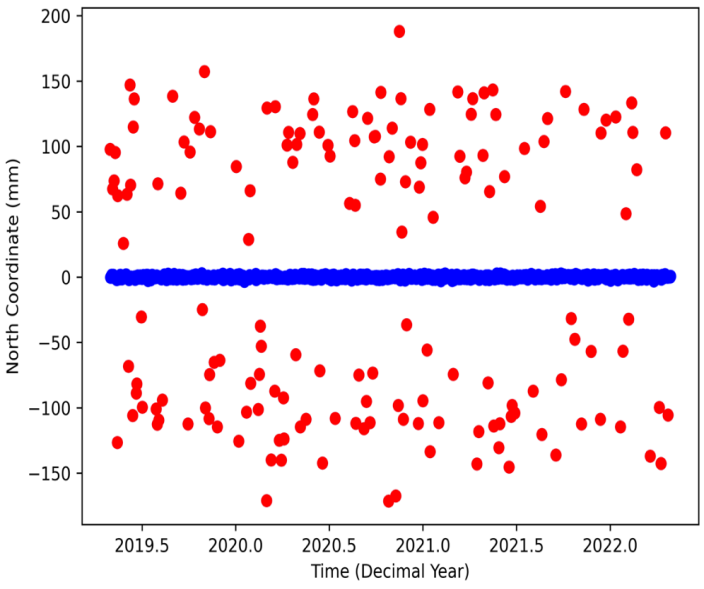
Để đánh giá hiệu quả của các phương pháp học máy LOF, O-C SVM và Isolation Forest trong phát hiện ngoại lai trong chuỗi tọa độ GNSS, nhóm nghiên cứu thực hiện theo sơ đồ như Hình 2.

Các bước thực hiện gồm:

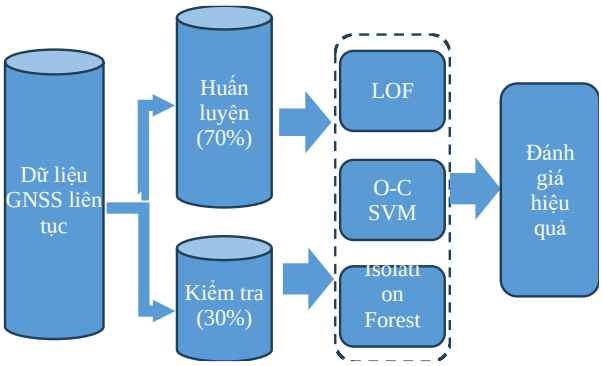
Bước 1: Chuẩn bị dữ liệu, từ dữ liệu chuỗi tọa độ GNSS của điểm SYNT được tạo thành, gán nhãn cho các điểm bình thường (không phải ngoại lai) là "0" và cho các điểm ngoại lai là "1".

Bước 2: Chia tập dữ liệu thành tập huấn luyện (70%) và tập kiểm tra (30%).

Bước 3: Huấn luyện mô hình LOF, O-C SVM và Isolation Forest trên tập huấn luyện. Xác định



Hình 1. Chuỗi toạ độ GNSS của điểm SYNT (Trong đó: North Coordinate – Tọa độ North; Time (Decimal Year) – Thời gian (decimet năm)).



Hình 2. Các bước thực hiện đánh giá hiệu quả các phương pháp máy học.

ngoại lai bằng cách sử dụng mô hình đã được huấn luyện với tập kiểm tra.

Bước 4: Đánh giá hiệu quả các mô hình.

Ở bước 1, việc gán nhãn đối với dữ liệu là nhằm để định danh ngoại lai trong quá trình huấn luyện mô hình (với tập dữ liệu huấn luyện) và chỉ sử dụng để so sánh kiểm tra sau khi huấn luyện (với tập dữ liệu kiểm tra). Do vậy, sau khi huấn luyện, với dữ liệu thực tế, các mô hình vẫn phát hiện được ngoại lai hiệu quả như vốn có mà không cần các nhãn này.

Ở bước 4, việc đánh giá hiệu quả của các phương pháp (mô hình) thông qua các tham số sau (Breunig và nnk., 2000):

Tỉ lệ số điểm ngoại lai được phát hiện chính xác:

$$\text{Precission} = \frac{TP}{TP+FP} \quad (5)$$

Tỉ lệ số điểm ngoại lai được tìm thấy:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Trung bình điều hòa giữa Precision và Recall:

$$F1\_score = 2\frac{Precission.Recall}{Precission + Recall} \quad (7)$$

Trong đó:

TP (True Positives), FP (False Positives) tương ứng là số lượng dữ liệu ngoại lai được phân loại đúng, sai vào lớp dữ liệu ngoại lai.

TN (True Negatives), FN (False Negatives) tương ứng là số lượng dữ liệu bình thường được phân loại đúng và sai vào lớp dữ liệu bình thường.

Trong nghiên cứu này, đường cong ROC (Receiver Operating Characteristic) và diện tích dưới đường cong UAC (Area Under the Curve) (Bradley, 1997) cũng được sử dụng trong phân loại và đánh giá hiệu suất của các phương pháp.

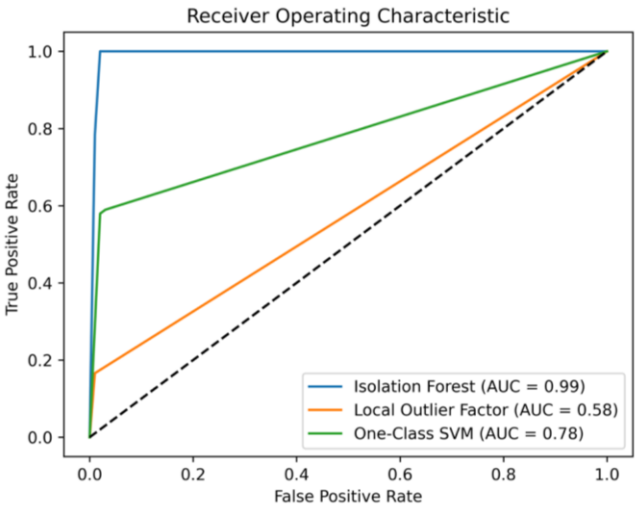
4. Kết quả nghiên cứu và thảo luận

Sau khi thực hiện các thuật toán để xây dựng mô hình và phát hiện các ngoại lai trong chuỗi tọa độ GNSS của điểm SYNT sử dụng ngôn ngữ lập trình Python (Van Rossum & Drake, 2003) trên cùng một nền tảng phần cứng. Kết quả huấn luyện, kiểm tra các mô hình được thể hiện trong Bảng 1.

Bảng 1 cho thấy phương pháp Isolation Forest có độ chính xác cao trong việc xác định các điểm dữ liệu bình thường (lớp 0) và ngoại lai (lớp 1) với Precission là 1,00 và 0,93 tương ứng. Phương pháp này có khả năng tìm ra hầu hết các điểm dữ liệu bình thường và ngoại lai với Recall tương ứng là 0,99 và 1,00. Với F1-score là 0,99 cho lớp 0 và 0,96 cho lớp 1, cho thấy hiệu suất tốt của mô hình trong cả hai lớp. Phương pháp LOF có tham số Precision là 0,87 cho lớp 0 và 0,77 cho lớp 1, cho thấy độ chính xác thấp hơn so với Isolation Forest trong việc xác định các điểm dữ liệu. Tham số Recall là 0,99 cho lớp 0 và 0,16 cho lớp 1, cho thấy phương pháp này có khả năng tìm ra hầu hết các điểm dữ liệu bình thường, nhưng lại không có khả năng trong việc phát hiện các điểm dữ liệu ngoại lai. Phương pháp O-C SVM có Precision là 0,93 cho lớp 0 và 0,83 cho lớp 1, cho thấy có độ chính xác cao trong việc xác định các điểm dữ liệu bình thường, nhưng kém hơn một chút trong việc phát hiện các điểm dữ liệu ngoại lai. Tham số Recall là 0,98 cho lớp 0 và 0,59 cho lớp 1, cho thấy phương pháp có khả năng tìm ra hầu hết các điểm dữ liệu bình thường, nhưng hiệu suất phát hiện ngoại lai (lớp 1) không cao. Như vậy, tham số F1-score là 0,95 cho lớp 0 và 0,69 cho lớp 1, cho thấy phương pháp có hiệu suất tốt trong việc xác định các điểm dữ liệu bình thường (lớp 0), nhưng kém hơn trong việc phát hiện các điểm dữ liệu ngoại lai (lớp 1).

Bảng 1. Kết quả đánh giá của các phương pháp trong phát hiện ngoại lai của chuỗi toạ độ GNSS.

Phương pháp	Lớp	Precision	Recall	F1-score
Isolation Forest	0	1,00	0,99	0,99
	1	0,93	1,00	0,96
LOF	0	0,87	0,99	0,93
	1	0,77	0,16	0,27
O-M SVM	0	0,93	0,98	0,95
	1	0,83	0,59	0,69



Hình 3. Đường cong ROC của các phương pháp trong phát hiện ngoại lai của chuỗi toạ độ GNSS điểm SYNC (Trong đó: Receiver Operating Characteristic - Tính chất hoạt động của máy thu; True Positive Rate - Tỷ lệ vị trí thực; False Positive Rate - Tỷ lệ sai số vị trí).

Kết quả tính đường cong ROC và diện tích dưới đường cong AUC được thể hiện trong Hình 3.

Hình 3 cho thấy, phương pháp Isolation Forest có diện tích dưới đường cong ROC (AUC) là 0,99, gần với 1, cho thấy mô hình có khả năng phân loại tốt giữa các điểm dữ liệu bình thường và ngoại lai. Phương pháp O-C SVM kém hơn, có AUC là 0,78, tương đối lớn, cho thấy mô hình có khả năng phân loại tương đối tốt giữa các điểm dữ liệu bình thường và ngoại lai, nhưng chưa đạt được mức đánh giá cao như Isolation Forest. Trong khi đó, phương pháp LOF có khả năng phân loại kém nhất giữa các điểm dữ liệu bình thường và ngoại lai với AUC thấp (0,58).

Như vậy, với số liệu chuỗi tọa độ điểm GNSS của điểm SYNT, kết quả phát hiện ngoại lai trong chuỗi bằng các phương pháp học máy cho thấy phương pháp Isolation Forest có hiệu suất tốt nhất, với độ chính xác cao, khả năng tìm ra hầu hết các điểm dữ liệu ngoại lai và diện tích dưới đường cong ROC (AUC) gần với 1. Local Outlier Factor có hiệu suất kém hơn, đặc biệt trong việc phát hiện các điểm dữ liệu ngoại lai. O-C SVM có độ chính xác tương đối cao trong việc xác định các điểm dữ liệu bình thường, nhưng có hiệu suất kém hơn trong việc phát hiện ngoại lai.

5. Kết luận

Các ngoại lai xuất hiện trong chuỗi tọa độ GNSS là hiện tượng phổ biến, chúng làm giảm chất lượng, độ chính xác của chuỗi tọa độ, việc loại bỏ ngoại lai là cần thiết khi xử lý số liệu. Trong nghiên cứu này, nhóm nghiên cứu đã tạo chuỗi tọa độ GNSS của điểm SYNT, với tỷ lệ ngoại lai là 15%, một tỷ lệ lớn mà các phương pháp thông thường không hiệu quả, và phát hiện ngoại lai trong chuỗi này bằng cách sử dụng ba phương pháp học máy: Isolation Forest, LOF và O-C SVM. Kết quả nghiên cứu cho thấy:

Phương pháp Isolation Forest đạt được hiệu suất tốt nhất trong việc phát hiện ngoại lai. Mô hình này có độ chính xác cao, khả năng tìm ra hầu hết các điểm dữ liệu ngoại lai và diện tích dưới đường cong ROC (AUC) gần với 1. Các chỉ số Precision, Recall và F1-score đều cho thấy hiệu suất tốt của mô hình trong cả hai lớp dữ liệu ngoại lai và dữ liệu bình thường. Phương pháp Local Outlier Factor cho thấy hiệu suất kém hơn so với Isolation Forest, đặc biệt là trong việc phát hiện các điểm dữ liệu ngoại lai. Mô hình này có độ chính xác thấp và không có khả năng phát hiện nhiều điểm dữ liệu ngoại lai. Phương pháp One-Class SVM có độ chính xác tương đối cao trong việc xác định các điểm dữ liệu bình thường, nhưng hiệu suất phát hiện ngoại lai không cao. Mô hình này có khả năng tìm ra hầu hết các điểm dữ liệu bình thường, nhưng vẫn còn kém trong việc phát hiện các điểm dữ liệu ngoại lai.

Dựa trên kết quả đánh giá và so sánh, phương pháp Isolation Forest là lựa chọn tốt nhất trong việc phát hiện ngoại lai trong chuỗi tọa độ GNSS của điểm SYNT, nó đạt được hiệu suất tốt, độ chính xác cao và khả năng phát hiện nhiều điểm dữ liệu ngoại lai.

Trong hướng nghiên cứu tiếp theo, nhóm nghiên cứu hướng tới nâng cao hiệu suất và độ chính xác của các phương pháp máy học trong phát hiện ngoại lai trong chuỗi tọa độ GNSS, trên cơ sở áp dụng các kỹ thuật mới, phát triển các phương pháp kết hợp khác nhau để tận dụng ưu điểm mỗi phương pháp.

Đóng góp của tác giả

Trần Đình Trọng - Đưa ra ý tưởng nghiên cứu và kiểm tra kết quả tính toán, hoàn thiện bản thảo; Nguyễn Đình Huy - Thực hiện tính toán trên cơ sở lý thuyết, lập trình Python, viết bản thảo.

Lời cám ơn

Nhóm tác giả xin trân trọng cảm ơn Trường Đại học Xây dựng Hà Nội đã giúp đỡ thực hiện nghiên cứu này thông qua đề tài khoa học công nghệ mã số XDA2022-05.

Tài liệu tham khảo

Bevis, M., Jonathan, B., & Dana J., C. I. I., (2020). The Art and Science of Trajectory Modelling. In J.-P. Montillet & M. S. Bos (Eds.), *Geodetic Time Series Analysis in Earth Sciences* (1st ed., pp. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-030-21718-1

Bradley, A. P., (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J., (2000). LOF. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104. https://doi.org/10.1145/342009.335388.

Dinh, T. T., Nguyen, D. H., Vu, N. Q., & long Nguyen, Q. (2023). Crustal displacement in Vietnam using CORS data during 2018-2021. *Earth Sciences Research Journal*, *27*(1), 27-36. https://doi.org/10.15446/esrj.v27n1.102630

Gao, W., Li, Z., Chen, Q., Jiang, W., & Feng, Y., (2022). Modelling and prediction of GNSS time series using GBDT, LSTM and SVM machine learning approaches. *Journal of Geodesy*, *96*(10), 71. https://doi.org/10.1007/s00190-022-01662-5.

Gomo, S., Durrheim, R. J., & Cooper, G. R. J., (2017). *Analysis of GPS Position Time Series in Africa*.

Hawkins, D. M., (1980). *Identification of Outliers* (1st ed.). Springer Netherlands. https://doi.org/10.1007/978-94-015-3994-4.

Hieu, H. T., Chou, T. Y., Fang, Y. M., & Hoang, T. V., (2018). Statistical process control methods for detecting outliers in GPS time series data. *Int Refereed J Eng Sci*, *7*(5), 8–15.

Ito, C., Takahashi, H., & Ohzono, M., (2019). Estimation of convergence boundary location and velocity between tectonic plates in northern Hokkaido inferred by GNSS velocity data. *Earth, Planets and Space*, *71*(1), 86. https://doi.org/10.1186/s40623-019-1065-z.

Kall, T., Oja, T., Kruusla, K., & Liibusk, A., (2021). New 3D velocity model of Estonia from GNSS measurements. *Estonian Journal of Earth Sciences*, *70*(2), 107. https://doi.org/10.3176/earth.2021.08.

Kiani, M., (2020). A specifically designed machine learning algorithm for GNSS position time series prediction and its applications in outlier and anomaly detection and earthquake prediction. *ArXiv Preprint ArXiv:2006.09067*.

Klos, A., Bogusz, J. B., Bos, M. S., & Gruszczynska, M., (2020). Modelling the GNSS Time Series: Different Approaches to Extract Seasonal Signals. In J.-P. Montillet & M. S. Bos (Eds.), *Geodetic Time Series Analysis in Earth Sciences* (1st ed., pp. 2–4). Springer International Publishing. https://doi.org/10.1007/978-3-030-21718-1.

Liu, F. T., Ting, K. M., & Zhou, Z.-H., (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17.

Métivier, L., Collilieux, X., Lercier, D., Altamimi, Z., & Beauducel, F., (2014). Global coseismic deformations, GNSS time series analysis, and earthquake scaling laws. *Journal of Geophysical Research: Solid Earth*, *119*(12), 9095–9109. https://doi.org/10.1002/2014jb011280.

Montillet, J.-P., & Bos, M. S., (2020). *Geodetic Time Series Analysis in Earth Sciences* (J.-P. Montillet & M. S. Bos, Eds.; 1st ed.). Springer International Publishing. https://doi.org/10.1007/978-3-030-21718-1.

Montillet, J.-P., Williams, S. D. P., Koulali, A., & McClusky, S. C., (2015). Estimation of offsets in GPS time-series and application to the detection of earthquake deformation in the far-field. *Geophysical Journal International*, *200*(2), 1207–1221. https://doi.org/10.1093/gji/ggu473.

Phong, D. V., Trọng, N. G., Chiến, N. V., Thành, N. H., Hà, L. L., Quân, N. V., & Quang, P. N. (2023). Phân tích chuyển dịch thẳng đứng vỏ Trái đất sử dụng hàm ANN từ kết quả xử lý chuỗi dữ liệu GNSS theo thời gian. *Tạp Chí Khí Tượng Thủy Văn*, *752*, 41–50. https://doi.org/10.36335/VNJHM.2022(752).41-50.

Riel, B., Simons, M., Agram, P., & Zhan, Z., (2014). Detecting transient signals in geodetic time series using sparse estimation techniques. *Journal of Geophysical Research: Solid Earth*, *119*(6), 5140–5160. https://doi.org/10.1002/2014JB011077.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C., (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, *13*(7), 1443–1471. https://doi.org/10.1162/089976601750264965.

Teferle, F. N., Williams, S. D. P., Kierulf, H. P., Bingley, R. M., & Plag, H.-P., (2008). A continuous GPS coordinate time series analysis strategy for high-accuracy vertical land movements. *Physics and Chemistry of the Earth, Parts A/B/C*, *33*(3–4), 205–216. https://doi.org/10.1016/j.pce.2006.11.002.

Tran, D. T. (2013). *Analyse rapide et robuste des solutions GPS pour la tectonique* [Université de Nice Sophia - Antipolis]. https://www.theses.fr/2013NICE4033.

Tran, D. T., Nguyen, Q. L., & Nguyen, D. H., (2021). General Geometric Model of GNSS Position Time Series for Crustal Deformation Studies -A Case Study of CORS Stations in Vietnam. *Journal of the Polish Mineral Engineering Society*, *1*(2), 183–198. https://doi.org/10.29227/IM-2021-02-16.

Tran, D. T., Nocquet, J.-M., Luong, N. D., & Nguyen, D. H., (2022). Determination of Helmert transformation parameters for continuous GNSS networks: a case study of the Géoazur GNSS network. *Geo-Spatial Information Science*, 1–14. https://doi.org/10.1080/10095020.2022.2138569.

Trần, Đ. T., Vũ, Đ. C., & Đào, D. T., (2014). Phương pháp Dikin phát hiện trị đo chứa sai số thô. *Tạp Chí Khoa Học và Công Nghệ, Viện Hàn Lâm Khoa Học Việt Nam*, *52*(4B), 519–526. https://www.researchgate.net/publication/277599309_Phuong_phap_Dikin_phat_hien_tri_do_chua_sai_so_tho.

Tran, T. D., Dao, T. D., Vu, T. S., Luong, D. N., Vu, C. D., Bui, S. N., & Ha, H. T., (2016). Outlier detection in GNSS position time series. *Science and Technology Development Journal*, *19*(2), 43–50. https://doi.org/10.32508/stdj.v19i2.665

Trọng, T. Đ., & Huy, N. Đ. (2023). Nghiên cứu xác định thời gian tắt dần sau động đất trong chuỗi tọa độ GNSS liên tục. *Tạp chí Khoa học Công nghệ Xây dựng (KHCNXD)-ĐHXDHN*. https://stce.huce.edu.vn/index.php/vn/article/view/2659.

Tsai, M.-C., Yu, S.-B., Shin, T.-C., Kuo, K.-W., Leu, P.-L., Chang, C.-H., & Ho, M.-Y. (2015). Velocity Field Derived from Taiwan Continuous GPS Array (2007 - 2013). *Terrestrial, Atmospheric and Oceanic Sciences*, *26*(5), 527. https://doi.org/10.3319/TAO.2015.05.21.01(T).

Van Rossum, G., & Drake, F. L. (2003). *An introduction to Python*. Network Theory Ltd. Bristol.

Wu, D., Yan, H., & Shen, Y., (2017). TSAnalyzer, a GNSS time series analysis software. *GPS Solutions*, *21*(3), 1389–1394. https://doi.org/10.1007/s10291-017-0637-2.