Đánh giá hiệu suất của các thuật toán học máy để phát hiện địa chấn

Trần Thành Công*, Đào Tăng Tín
Trường Đại học Quốc tế Hồng Bàng

TÓM TẮT
Các mối nguy hiểm của hoạt động khai thác mỏ rất khó phát hiện, có thể được so sánh với các trận động đất và gây ra hậu quả rất nghiêm trọng cho con người. Do đó, phát triển các phương pháp dự đoán trạng thái nguy hiểm dưới hầm mỏ là cần thiết để giảm thiểu những thiệt hại không mong muốn. Trong bài báo này, các thuật toán học máy như k láng giếng gần nhất, cây quyết định, và RUSBoost được áp dụng lên tập dữ liệu để phát hiện ra các trạng thái nguy hiểm. Dựa trên độ chính xác phân loại, tỷ lệ dương thực sự, giá trị dự đoán tích cực, và độ đo F1, chúng tôi nhận thấy rằng các thuật toán học máy đã được đề cập có thể tạo ra các mô hình phân loại với kết quả đáng tin cậy.

Từ khoá: *thuật toán học máy, k láng giếng gần nhất, cây quyết định, RUSBoost, địa chấn*

1. GIỚI THIỆU

Hoạt động khai thác mỏ gắn liền với sự xuất hiện của các mối nguy hiểm. Nguy cơ địa chấn thường xuyên xảy ra ở nhiều hầm mỏ được xem là trường hợp đặc biệt của mối nguy hiểm này. Hiểm họa địa chấn là nguy cơ khó phát hiện, có thể dự đoán được trong các nguy cơ tự nhiên và có thể so sánh với một trận động đất. Có rất nhiều hệ thống giám sát địa chấn tiên tiến và hiện đại cho phép hiểu rõ hơn các quá trình khối đá và định nghĩa các phương pháp dự báo nguy cơ địa chấn. Tuy nhiên, độ chính xác của các phương pháp được tạo ra vẫn chưa hoàn hảo. Sự phức tạp của các quá trình địa chấn và sự chênh lệch lớn giữa số lượng các sự kiện địa chấn năng lượng thấp và số hiện tượng năng lượng cao khiến các kỹ thuật thống kê không đủ để dự đoán nguy cơ địa chấn. Do đó, điều cần thiết là phải tìm kiếm các phương án mới để dự đoán nguy cơ tốt hơn.

Thời gian gần đây, các phương pháp học máy đã thu hút được nhiều sự quan tâm trong cộng đồng nghiên cứu [1]. Như được mô tả trong nhiều nghiên cứu trong thời gian gần đây, các kỹ thuật học máy có triển vọng mang lại độ chính xác cao trong phân loại. Việc đạt được độ chính xác cao trong dự đoán là rất quan trọng vì nó có thể mang lại các biện pháp bảo vệ thích hợp trong các ứng dụng cụ thể. Trong các đánh giá nguy cơ địa chấn, kỹ thuật phân nhóm dữ liệu đã áp dụng[2], và để dự đoán chấn động địa chấn, mạng nơron nhân tạo cũng được sử dụng [3]. Trong phần lớn các ứng dụng, kết quả thu được bằng các phương pháp này đã được báo cáo dưới dạng hai trạng thái, được hiểu là trạng thái nguy hiểm và trạng thái không nguy hiểm. Dữ liệu phân phối không cân bằng giữa trạng thái nguy hiểm và trạng thái không nguy hiểm là một vấn đề nghiêm trọng trong dự báo nguy cơ địa chấn. Tuy nhiên, các phương pháp được sử dụng hiện nay vẫn chưa đủ để đạt được độ nhạy và độ đặc hiệu tốt của các dự đoán.

Bên cạnh đó, trong bài báo của Bogucki [4], nhóm tác giả đã sử dụng các mô hình học máy khác nhau, như phân tích phân biệt tuyến tính

Tác giả liên hệ: ThS. Trần Thành Công
Email: congtt@hiu.vn

(LDA), hồi quy logistic (LR), bộ phân loại cây phụ (ETC) và bộ phân loại tăng cường độ phân giải cực cao (XGB), để giải quyết vấn đề phân loại liệu tổng năng lượng địa chấn trong vài giờ tới có đạt đến mức cảnh báo hay không. Trong bài báo của Netti[5], các tác giả đã đề xuất một cách tiếp cận mới để cải thiện độ chính xác của việc dự đoán nguy cơ địa chấn bằng cách sử dụng bộ phân loại Naive Bayes với xử lý phủ định. Cách tiếp cận này đã cho thấy kết quả vượt trội hơn so với bộ phân loại Naïve Bayes truyền thống về độ chính xác. Để đánh giá toàn diện mô hình học máy phân loại địa chấn cần dựa vào các tiêu chí đánh giá phân loại khác nhau, tuy nhiên, các nghiên cứu trên chưa áp dụng đa dạng các tiêu chí đánh giá. Sự cần thiết phải sử dụng các tiêu chính đánh giá khác nhau bởi vì kết quả của các tiêu chí đánh giá hiệu suất phân loại của mô hình học máy là một trong những yếu tố quan trọng để chọn lựa mô hình học máy tốt nhất trong số các mô hình học máy khác nhau để áp dụng vào thực tế. Trong bài báo của Kurach[6], nhóm tác giả đã trình bày giải pháp cho thách thức khai thác dữ liệu AAIA'16 dựa trên mạng thần kinh tái diễn (RNN) với các ô nhớ ngắn hạn dài (LSTM). Kết quả phân loại các hoạt động địa chấn từ giải pháp trên thu được rất cao. Trong bài báo của Geng[7], nhóm tác giả đã góp phần giải quyết vấn đề phụ thuộc lịch sử lâu dài vào dự đoán chuỗi thời gian địa chấn với mạng nơron tích chập theo thời gian sâu (CNN). Nhóm tác giả này đã đề xuất một mạng tích chập theo thời gian nhân quả giãn nở (DCTCNN) và mô hình kết hợp bộ nhớ ngắn hạn dài hạn CNN (CNN - LSTM) để dự báo các sự kiện địa chấn. Tuy nhiên, các mô hình này đắt về mặt tính toán và hoạt động tốt hơn trên các tập dữ liệu lớn hơn[8].

Trong bài báo này, ba phương pháp học máy phổ biến nhưng rất hiệu quả trong phân loại, bao gồm phương pháp k láng giềng gần nhất (KNN), cây quyết định (DT), và RUSBoost được áp dụng lên tập dữ liệu để tạo ra các mô hình phân loại giữa trạng thái nguy hiểm và trạng thái không nguy hiểm. Sau đó, các tiêu chí như độ chính xác phân loại, tỷ lệ dương thực sự, giá trị dự đoán tích cực, và độ đo F1 được sử dụng để đánh giá hiệu suất của các mô hình thu được từ ba thuật toán học máy.

Bài báo này được chia làm 5 mục. Mục 1 là giới thiệu. Mục 2 miêu tả đặc điểm của dữ liệu. Mục 3 thảo luận về các thuật toán học máy và các tiêu chí đánh giá hiệu suất phân loại. Mục 4 nêu lên kết quả của các thuật toán học máy đã áp dụng vào dữ liệu đã được miêu tả ở mục 2. Phần kết luận chung của bài báo sẽ được thể hiện ở mục 5.

2. ĐẶC ĐIỂM CỦA DỮ LIỆU

Dữ liệu sử dụng trong bài báo này được trích dẫn từ tài liệu số 9. Tập dữ liệu trình bày có đặc trưng bởi sự phân bố không cân bằng giữa trạng thái nguy hiểm và trạng thái không nguy hiểm. Trong tập dữ liệu này bao gồm 2584 mẫu, nhưng chỉ có 170 mẫu trạng thái nguy hiểm. Trong bảng dữ liệu có tất cả 18 cột thuộc tính (cột 1 - cột 18), và một cột nhãn (cột 19). Các thông tin thuộc tính được trình bày cụ thể như sau:

(1) seismic: kết quả đánh giá nguy cơ địa chấn chuyển dịch trong mỏ làm việc theo phương pháp địa chấn (a - thiếu nguy hiểm, b - nguy hiểm thấp, c - nguy hiểm cao, d - trạng thái nguy hiểm);

(2) seismoacoustic: kết quả đánh giá nguy cơ địa chấn dịch chuyển trong mỏ làm việc theo phương pháp địa âm học;

(3) shift: thông tin về loại ca (W - ca nhận than, N - ca làm việc);

(4) genergy: năng lượng địa chấn được ghi lại trong ca trước bằng thiết bị đo địa chất hoạt động mạnh nhất (GMax) trong số các thiết bị đo địa chất;

(5) gpuls: một số xung được GMax ghi lại trong ca trước;

(6) gdenergy: độ lệch của năng lượng được GMax ghi nhận trong ca trước với năng lượng trung bình được ghi nhận trong tám ca trước;

(7) gdpuls: độ lệch của một số xung được ghi nhận trong ca trước bởi GMax so với số xung trung bình được ghi lại trong tám ca trước;

(8) ghazard: kết quả đánh giá nguy cơ địa chấn dịch chuyển trong mỏ đang làm việc

thu được bằng phương pháp địa âm học chỉ dựa trên mẫu đăng ký GMax;

(9) nbumps: số lượng các va chạm địa chấn được ghi lại trong ca trước;

(10) nbumps2: số lần va chạm địa chấn (trong dải năng lượng [10^2, 10^3]) được đăng ký trong ca trước;

(11) nbumps3: số lần va chạm địa chấn (trong khoảng năng lượng [10^3, 10^4]) được đăng ký trong ca trước;

(12) nbumps4: số lần va chạm địa chấn (trong phạm vi năng lượng [10^4, 10^5]) được đăng ký trong ca trước;

(13) nbumps5: số lần va chạm địa chấn (trong phạm vi năng lượng [10^5, 10^6]) được đăng ký trong ca trước;

(14) nbumps6: số lần va chạm địa chấn (trong phạm vi năng lượng [10^6, 10^7]) được đăng ký trong ca trước;

(15) nbumps7: số lần va chạm địa chấn (trong phạm vi năng lượng [10^7, 10^8]) được đăng ký trong ca trước;

(16) nbumps89: số lần va chạm địa chấn (trong phạm vi năng lượng [10^8, 10^9]) được đăng ký trong ca trước;

(17) energy: tổng năng lượng của các va chạm địa chấn đã đăng ký trong ca trước;

(18) maxenergy: năng lượng tối đa của các va chạm địa chấn được đăng ký trong ca trước;

(19) class: nhãn, '1' có nghĩa là va chạm địa chấn năng lượng cao xảy ra trong ca tiếp theo ('trạng thái nguy hiểm'), '0' có nghĩa là không có va chạm địa chấn năng lượng cao nào xảy ra trong ca tiếp theo ('trạng thái không nguy hiểm').

3. CÁC PHƯƠNG PHÁP HỌC MÁY

3.1. Thuật toán k láng giềng gần nhất (KNN)

Thuật toán KNN là một phương pháp phi tham số được sử dụng để giải quyết cả vấn đề phân loại và hồi quy. Mặc dù thuật toán này đơn giản, nhưng nó có thể phân loại tốt hơn nhiều thuật toán phức tạp hơn. KNN yêu cầu chỉ chọn k, số lượng hàng xóm được xem xét khi tiến hành phân loại. Nếu giá trị k nhỏ, ước lượng phân loại dễ bị sai số thống kê lớn. Ngược lại, nếu giá trị k lớn, điều này cho phép các điểm ở xa góp phần vào việc phân loại, gây ra sự bỏ qua một số chi tiết của phân bố lớp. Do đó, giá trị của k được xem xét để chọn nhằm giảm thiểu lỗi phân loại trên một số dữ liệu xác nhận độc lập hoặc bằng các thủ tục xác nhận chéo[10 - 11].

3.2. Thuật toán cây quyết định (Decision Tree - DT)

Thuật toán DT là một trong những thuật toán phân loại được sử dụng phổ biến nhất trong các bài toán phân loại của học máy. Điều này chủ yếu là do thuật toán này có khả năng xử lý các vấn đề phức tạp bằng cách cung cấp một biểu diễn dễ hiểu, dễ giải thích và cũng như khả năng thích ứng của chúng đối với nhiệm vụ suy luận bằng cách tạo ra các quy tắc phân loại logic. Thuật toán cây quyết định bao gồm các nút để kiểm tra các thuộc tính, các cạnh để phân nhánh theo các giá trị của thuộc tính đã chọn và để lại các lớp gắn nhãn trong đó đối với mỗi lá, một lớp duy nhất được đính kèm. Có hai thủ tục chính trong thuật toán cây quyết định cho thấy một thủ tục là để xây dựng cây, thủ tục kia cho phép suy luận tri thức, tức là phân loại [12 - 13].

3.3. Thuật toán RUSBoost

Một thuật toán lấy mẫu/ tăng cường dữ liệu kết hợp mới được gọi là RUSBoost, được thiết kế để cải thiện hiệu suất của các mô hình được đào tạo dựa trên dữ liệu sai lệch. RUSBoost áp dụng lấy mẫu dưới ngẫu nhiên. Thuật toán này loại bỏ ngẫu nhiên các ví dụ khỏi lớp đa số. **Hình 1** trình bày thuật toán RUSBoost. Trong bước 1, trọng số của mỗi ví dụ được khởi tạo là *1/m*, trong đó *m* là số ví dụ trong tập dữ liệu huấn luyện. Trong bước 2, giả thuyết yếu tố được huấn luyện lặp đi lặp lại. Trong bước 2a, lấy mẫu dưới ngẫu nhiên được áp dụng để loại bỏ các ví dụ về lớp đa số cho đến khi N% của tập dữ liệu huấn luyện mới (tạm thời), S_t' thuộc về lớp thiểu số. Do đó, S_t' sẽ có phân phối trọng lượng mới, D_t'. Trong bước 2b, S_t' và D_t' được chuyển cho người học cơ sở,

WeakLearn, tạo ra giả thuyết yếu, h_t ở bước 2c. Dựa trên tập dữ liệu đào tạo ban đầu, S và phân phối trọng lượng, D_t tổn thất giả, ϵ_t được tính ở bước 2d. Trong bước 2e, tham số cập nhật trọng lượng, α được tính bằng $\epsilon_t/(1-\epsilon_t)$. Tiếp theo, phân phối trọng lượng cho lần lặp tiếp theo, $D_{(t+1)}$ được cập nhật ở bước 2f và chuẩn hóa ở bước 2g. Sau T lần lặp của bước 2, giả thuyết cuối cùng, $H(x)$ được trả về dưới dạng biểu quyết có trọng số của các giả thuyết yếu [14].

Thuật toán RUSBoost

Cho: Tập hợp S các ví dụ $(x_1, y_1), \ldots, (x_m, y_m)$ với lớp thiểu số $y^r \in Y, |Y| = 2$

Người học yếu, WeakLearn

Số lần lặp lại, T

Tỷ lệ phần trăm mong muốn trong tổng số các trường hợp được đại diện bởi lớp thiểu số, N

1. Khởi tạo $D_1(i) = \frac{1}{m}$ với mọi i.
2. Gán t = 1, 2, ..., T
   a. Tạo tập dữ liệu huấn luyện tạm thời S'_t với phân phối D'_t bằng cách sử dụng lấy mẫu dưới ngẫu nhiên.
   b. Gọi WeakLearn, cung cấp cho nó các ví dụ S'_t và trọng lượng của chúng D'_t.
   c. Lấy lại một giả thuyết $h_t: X \times Y \to [0, 1]$.
   d. Tính tổn thất giả (đối với S và D_t):
   $\epsilon_t = \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$.
   e. Tính thông số cập nhật trọng lượng:
   $\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}$
   f. Cập nhật D_t:
   $D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i))}$
   g. Chuẩn hóa D_{t+1}: Cho $Z_t = \sum_i D_{t+1}(i)$
   $$D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}$$
3. Đưa ra giả thuyết cuối cùng:
$$H(x) = \underset{y \in Y}{argmax} \sum_{t=1}^{T} h_t(x, y) log \frac{1}{\alpha_t}$$

Hình 1. *Thuật toán RUSBoost*

3.4. Phương pháp đánh giá hiệu suất phân loại

Trong bài báo này, các tiêu chí như độ chính xác phân loại, tỷ lệ dương thực sự, giá trị dự đoán tích cực, và độ đo F1 được sử dụng để đánh giá hiệu suất của các thuật toán học máy nói trên dựa trên tập dữ liệu được mô tả ở mục 2. Trong phân loại nhị phân, ma trận nhầm lẫn được trình bày trong **Bảng 1** là một trong những cách tốt nhất để đánh giá các mô hình học máy. Ma trận nhầm lẫn cho thấy mối quan hệ giữa đầu ra của bộ phân loại và đầu ra đúng. **Bảng 1** chỉ ra bốn kết hợp khác nhau của các giá trị dự đoán và giá trị thực tế, được giải thích là dương tính thực (TP), dương tính giả (FP), âm tính thực sự (TN), âm tính giả (FN). Trong nghiên cứu này, các trạng thái nguy hiểm được xác định là loại tích cực, kí hiệu là 1 và các trạng thái không nguy hiểm được coi là loại tiêu cực, kí hiệu là 0. Dựa trên ma trận nhầm lẫn, độ chính xác phân loại, tỷ lệ dương thực sự, giá trị dự đoán tích cực, và độ đo F1 được tính toán [15].

Bảng 1. *Ma trận nhầm lẫn*

Giá trị thực tế	*Giá trị dự đoán*		
	Lớp	Tiêu cực (0)	Tích cực (1)
	Tiêu cực (0)	Âm tính thực sự (TN)	Dương tính giả (FP)
	Tích cực (1)	Âm tính giả (FN)	Dương tính thực sự (TP)

Các tiêu chính đánh giá hiệu suất mô hình học máy được định nghĩa như sau [16]:

Độ chính xác của phân loại là một trong những thước đo được sử dụng phổ biến nhất để đánh giá độ chính xác phân loại trong học máy. Độ chính xác phân loại được định nghĩa là tỷ lệ của các mẫu được phân loại chính xác trên tổng số mẫu và được trình bày như (1).

$$\text{Độ chính xác} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Giá trị dự đoán tích cực, P minh họa tỷ lệ các mẫu dương tính được phân loại chính xác trên tổng số các mẫu dự đoán dương tính, và được thể hiện như (2).

$$P = \frac{TP}{(TP + FP)} \quad (2)$$

Tỷ lệ dương thực sự hay độ nhạy, S thể hiện tỷ lệ giữa mẫu dương tính tích cực trên tổng số mẫu dương tính tích cực và số mẫu âm tính giả được trình bày như (3).

$$S = \frac{TP}{(TP + FN)} \quad (3)$$

Độ đo F1 được định nghĩa là giá trị trung bình hài hòa giữa giá trị dự đoán tích cực, P và tỷ lệ dương thực sự hay độ nhạy, S được minh họa trong (4). Độ đo F1 có phạm vi từ 0 đến 1, có nghĩa là giá trị cao của nó cho thấy hiệu suất phân loại cao.

$$F1 = 2 * \frac{P * S}{P + S} \quad (4)$$

4. KẾT QUẢ

Các ma trận nhầm lẫn của thuật toán KNN, DT và RUSBoost dựa trên tập dữ liệu được trình bày lần lượt trong **Bảng 2, 3, 4**. Dựa vào các ma trận nhầm lẫn đó, bảng tổng hợp TP, FP, TN, và FN của thuật toán KNN, DT, và RUSBoost được thể hiện ở **Bảng 5**. **Hình 2** so sánh các tỉ lệ TP, FP, TN và FN của thuật toán KNN, DT và RUSBoost. Thông qua **Hình 2**, chúng tôi nhận thấy thuật toán KNN dự đoán số lần âm tính thực sự cao nhất, tiếp theo sau lần lượt là thuật toán DT và RUSBoost.

Ngược lại, thuật toán RUSBoost dự đoán số lần dương tính thực sự là cao nhất, tiếp theo sau là thuật toán DT và KNN. Điều đáng chú ý là số lần dương tính thực sự của thuật toán KNN rất thấp. Bên cạnh đó, chúng tôi cũng nhận thấy tỷ lệ dương tính thực sự rất thấp hơn so với tỷ lệ âm tính thực sự ở cả ba thuật toán học máy nói trên.

Bảng 2. *Ma trận nhầm lẫn của thuật toán KNN*

Giá trị thực tế	Giá trị dự đoán		
	Lớp	Tiêu cực (0)	Tích cực (1)
	Tiêu cực (0)	716	7
	Tích cực (1)	52	1

Bảng 3. *Ma trận nhầm lẫn của thuật toán DT*

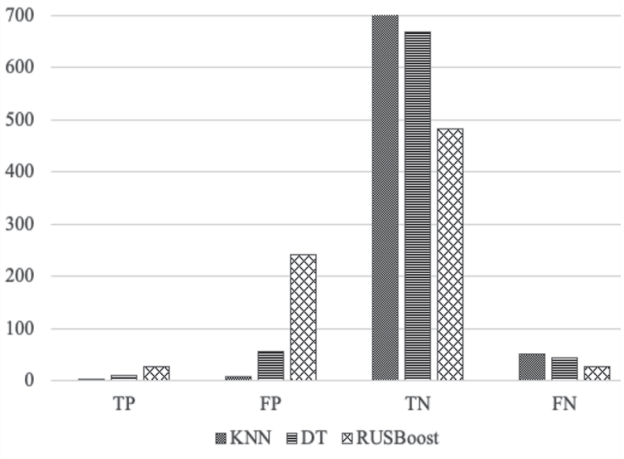
Giá trị thực tế	Giá trị dự toán		
	Lớp	Tiêu cực (0)	Tích cực (1)
	Tiêu cực (0)	668	55
	Tích cực (1)	44	9

Bảng 4. *Ma trận nhầm lẫn của thuật toán RUSBoost*

Giá trị thực tế	Giá trị dự đoán		
	Lớp	Tiêu cực (0)	Tích cực (1)
	Tiêu cực (0)	482	242
	Tích cực (1)	27	26

Bảng 5. *Bảng tổng hợp TP, FP, TN, và FN của thuật toán KNN, DT và RUSBoost*

	KNN	DT	RUSBoost
TP	1	9	26
FP	7	55	242
TN	716	668	482
FN	52	44	27



Hình 2. *So sánh 4 tỷ số TP, FP, TN, FN của thuật toán KNN, DT và RUSBoost*

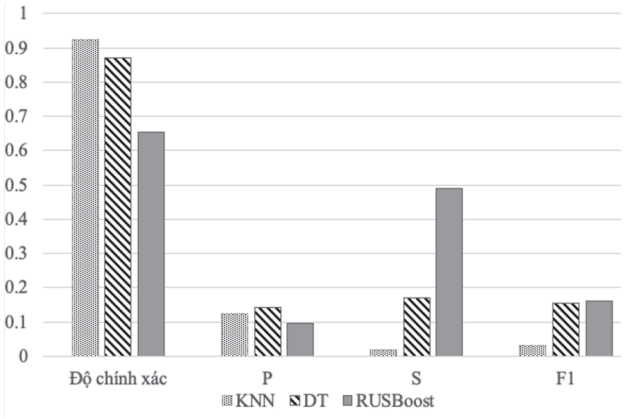
Bảng so sánh độ chính xác, P, S, và F1 của ba thuật toán KNN, DT, và RUSBoost được trình bày ở **Bảng 6**. Bên cạnh đó, các tiêu chí đo phân loại của thuật toán KNN, DT, và RUSBoost cũng được trình bày ở **Hình 3**. Dựa trên **Hình 3**, độ chính xác của loại thuật toán KNN là cao nhất, khoảng 0.924, tiếp đến là thuật toán DT và RUSBoost với độ chính xác lần lượt là 0.872 và 0.655. Trong khi đó, tỷ lệ dương thực sự cho thấy sự ngược lại, cao nhất là thuật toán RUSBoost, sau đó lần lượt là DT và KNN với các giá trị lần lượt là 0.49, 0.17 và 0.019.

Tương tự như vậy, độ đo F1 cũng cho thấy giá trị cao nhất ở thuật toán RUSBoost, khoảng 0.162, tiếp theo đó là thuật toán DT và KNN với các giá trị lần lượt là 0.153 và 0.032. Ở giá trị

dự đoán tích cực thì lại cho thấy kết quả cao nhất thuộc về thuật toán DT với giá trị là 0.141, tiếp theo sau là KNN và RUSBoost với các giá trị lần lượt là 0.125 và 0.097.

Bảng 6. *Bảng so sánh Độ chính xác P, S và F1 của ba thuật toán KNN, DT và RUSBoost*

	KNN	DT	RUSBoost
Độ chính xác	0.924	0.872	0.655
P	0.125	0.141	0.097
S	0.019	0.17	0.49
F1	0.032	0.153	0.162



Hình 3. *Các tiêu chí đo phân loại của thuật toán KNN, DT và RUSBoost*

5. KẾT LUẬN

Hoạt động khai thác mỏ luôn xuất hiện của các mối nguy hiểm dẫn đến các vụ tai nạn đáng tiếc xảy ra. Do đó, các phương pháp thích hợp cần được đề xuất để có thể giảm thiểu các vụ tai nạn không mong muốn xảy ra. Trong bài báo này, ba thuật toán học máy cơ bản, bao gồm KNN, DT và RUSBoost, được sử dụng để phân loại trạng thái nguy hiểm và trạng thái không nguy hiểm khi khai thác mỏ. Bài báo còn sử dụng đa dạng các tiêu chí đánh giá khác nhau, như độ chính xác phân loại, tỷ lệ dương thực sự, giá trị dự đoán tích cực, và độ đo F1, để đánh giá toàn diện hơn về mô hình phân loại địa chấn giữa hai trạng thái nguy hiểm và không nguy hiểm. Dựa trên tiêu chí độ chính xác, cả ba thuật toán nói trên đã cho thấy các kết quả phân loại tốt, trong đó nổi bật nhất là độ chính xác của thuật toán KNN với giá trị khoảng 0.924. Kết quả của các tiêu chí đánh giá khác như tỷ lệ dương thực sự, giá trị dự đoán tích cực, và độ đo F1 cũng thu được trong bài báo này. Tuy nhiên, các kết quả này vẫn chưa đạt tới các giá trị mong muốn.

Trong tương lai các kỹ thuật lấy mẫu lại được tìm hiểu và thực hiện để giúp chúng tôi giảm tỷ lệ mất cân bằng của hai trạng thái nguy hiểm và không nguy hiểm. Do đó, các kết quả phân loại có thể được cải thiện, đặc biệt là nâng cao tỷ lệ dương tính thực sự so với tỷ lệ âm tính thực sự.

TÀI LIỆU THAM KHẢO

[1] L. M. Bache K, "*UCI machine learning repository*," [Online]. Available: http://archive.ics.uci.edu/ml/index.php. [Accessed 20 9 2020].

[2] AndrzejLeśniak and ZbigniewIsakow, "*Space–time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland,*" *International Journal of Rock Mechanics and Mining Sciences,* vol. 46, no. 5, pp. 918-928, 2009.

[3] J. Kabiesz, "*Effect of the form of data on the quality of mine tremors hazard forecasting using neural networks,*"*Geotechnical and Geological Engineering, Springer ,* vol. 24, no. 5, p. 1131–1147, 2005.

[4] R. Bogucki, J. Lasek, J. K. Milczek and M. Tadeusiak, "*Early warning system for seismic events in Coal Mines using machine learning,*" in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Gdansk, 2016.

[5] K. Netti and Y. Radhika, "*An efficient Naïve Bayes classifier with negation handling for seismic hazard prediction,*" in *2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore*, Coimbatore, India, 2016.

[6] K. Kurach and K. Pawlowski, "*Predicting dangerous seismic activity with Recurrent*

Neural Networks," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS*, Gdansk, 2016.

[7] Y. Geng, L. Su, Y. Jia and C. Han, "*Seismic Events Prediction Using Deep Temporal Convolution Networks,*" *Journal of Electrical and Computer Engineering,* vol. 2019, pp. 1-14, 2019.

[8] "*Towards Data Science,*" [Online]. Available: https://towardsdatascience.com/deep-learning-vs-classical-machine-learning-9a42c6d48aa. [Accessed 27 10 2020].

[9] "*UCI, Machine Learning Repository,*" [Online]. Available: https://archive.ics.uci.edu/ml/datasets/seismic-bumps.

[10] O. Beckonert, M. E. Bollard, T. M. Ebbels, H. C. Keun, H. Antti, E. Holmes, J. C. Lindon and J. K. Nicholson, "*NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches,*" *Analytica Chimica Acta,* vol. 490, no. 1-2, p. 3–15, 2003.

[11] B. Alsbergav, R. Goodacrea, J. Rowlandb and D. Kella, "*Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods,*" *Analytica Chimica Acta,* vol. 348, no. 1-3, pp. 389-407, 1997.

[12] J. R. Quinlan, C4.5 : programs for machine learning, San Mateo, Calif. : Morgan Kaufmann Publishers, 1993.

[13] J. R. Quinlan, "*Induction of decision trees,*" *Machine Learning ,* vol. 1, p. 81–106, 1986.

[14] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse and A. Napolitano, "*RUSBoost: Improving Classification Performance when Training Data is Skewed,*" in *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA, 2008.

[15] S.Kanmania, V. P.Thambiduraia, V.Sankaranarayanan and P.Thambiduraia, "*Object-oriented software fault prediction using neural networks,*" *Information and Software Technology ,* vol. 49, no. 5, pp. 483-492, 2007.

[16] A. K. Dwivedi, "*Performance evaluation of different machine learning techniques for prediction of heart disease,*" *Neural Computing and Applications,* vol. 29, no. 10, p. 685–693, 2018.

Performance evaluation of machine learning Algorithms for predicting hazardous seismic bumps

Tran Thanh Cong*, Dao Tang Tin

ABSTRACT

Mining operations are always at risk because there are seismic hazards that frequently occur underneath mines. These hazards are difficult to detect, can be compared with earthquakes, and have very serious consequences for humans. Therefore, the development of methods to predict dangerous underground conditions is essential to minimize undesirable damage. In this paper, machine learning algorithms, such as k nearest neighbors, decision trees, and RUSBoost are applied to the dataset to detect dangerous and non - dangerous states. Based

on classification measurements, named accuracy, recall, precision, and F1 measure, we observe the machine learning algorithms mentioned can create the classification models with reliable results.

Keywords: *machine learning, KNN classification, decision tree, RUSBoost, seismic bump*

Received: 24/09/2020
Revised: 02/11/2020
Accepted for publication: 09/11/2020