*Tạp chí Phát triển Khoa học và Công nghệ – Engineering and Technology 2024, ():1-10*

Open Access Full Text Article

*Bài nghiên cứu*

# Ứng dụng học máy trong phân chia đơn nguyên địa chất công trình dựa trên dữ liệu thí nghiệm đất trong phòng

**Huỳnh Văn Thịnh[1], Kiều Lê Thủy Chung[2,3], Lê Minh Sơn[4], Ngô Tấn Phong[2,3,*]**

Use your smartphone to scan this QR code and download this article

**TÓM TẮT**

Đơn nguyên địa chất công trình (ĐCCT) là một thể tích đất đá đồng nhất có cùng tên gọi và các đặc trưng cơ lý biến thiên không có tính quy luật, hoặc nếu các đặc trưng cơ lý biến thiên có quy luật thì quy luật này có thể bỏ qua khi thỏa mãn điều kiện thống kê. Phân chia đơn nguyên ĐCCT là công việc rất quan trọng nhằm cung cấp dữ liệu tin cậy cho tính toán, thiết kế nền móng và lựa chọn giải pháp thi công hợp lý góp phần tiết kiệm chi phí khi xây dựng công trình. Hiện tại, việc phân chia đơn nguyên thường được tiến hành thủ công trên cơ sở các tiêu chuẩn phân loại đất của Việt Nam (TCVN), Mỹ (ASTM), Anh (BS)… Vì công việc được tiến hành thủ công nên không thể tránh khỏi các sai sót, đặc biệt khi tiến hành với bộ dữ liệu lớn. Nhằm giảm bớt các sai sót khi phân chia đơn nguyên cũng như phát hiện các dị thường trong cấu trúc ĐCCT và tiết kiệm thời gian khi tổng hợp lượng dữ liệu lớn để xây dựng cơ sở dữ liệu ĐCCT cho một khu vực, bài báo này ứng dụng học máy để phân chia đơn nguyên ĐCCT một cách tự động trên cơ sở sử dụng ba thuật toán phân cụm là K-Means, Gaussian Mixture Model (GMM) và Mean Shift. Bộ số liệu đầu vào được lấy từ kết quả thí nghiệm cơ lý của 437 mẫu đất ở khu vực Quận 1 và Quận 8 thuộc Tp. Hồ Chí Minh. Kết quả cho thấy, chương trình phân chia tự động cho kết quả phân chia các đơn nguyên ĐCCT rất trùng khớp với kết quả theo phương pháp thủ công. Hơn nữa, chương trình tự động có thể phân chia ra các đơn nguyên rất chi tiết, điều này giúp phát hiện ra các dị thường trong cấu trúc ĐCCT.

**Từ khoá:** Gaussian Mixture Model, Học máy, K-Means, Mean Shift, Phân loại đất

## GIỚI THIỆU

Đơn nguyên ĐCCT là một thể tích đất đá đồng nhất có cùng tên gọi và các đặc trưng cơ lý biến thiên không có tính quy luật, hoặc nếu các đặc trưng cơ lý biến thiên có quy luật thì quy luật này có thể bỏ qua khi thỏa mãn điều kiện thống kê. Hiện nay công tác phân loại và gọi tên đất ở Việt Nam còn được thực hiện khá thủ công. Một số nghiên cứu về phân loại đất tiêu biểu như Huỳnh Văn Bình, Lê Phước Hảo [1] nghiên cứu về hệ thống phân loại đất phục vụ cho xây dựng các công trình thủy điện khu vực Tây Nguyên; Thái Thành Dư, Ông Văn Ninh, Phạm Thanh Vũ và nnk [2] nghiên cứu về các đặc tính phân loại đất theo hệ thống phân loại WRB (World Reference Base) 2006 tỉnh Hậu Giang tỉ lệ 1/100.000; Võ Quang Minh, Lê Quang Trí [3] với đề tài đất Đồng bằng sông Cửu Long phân loại theo hệ thống WRB. Nhìn chung, hiện các nghiên cứu về phân loại đất được thực hiện chủ yếu dựa vào bản đồ địa tầng khu vực hay một hệ thống phân loại nào đó mà hầu như chưa áp dụng phương pháp học máy.

Trên thế giới, có nhiều nghiên cứu về phân loại đất sử ứng dụng phương pháp học máy như Xiangrong Wang, Hui Wang, Robert Y. Liang, Yang Liu [4] sử dụng phương pháp bán phân cụm có giám sát để phân loại đất dựa trên dữ liệu hố khoan và thí nghiệm xuyên tĩnh; Fausto Molina-Gómez, António Viana da Fonseca và nnk [5] ứng học máy để xác định sự phân bố các lớp đất dựa trên kết quả thí nghiệm xuyên tĩnh đo áp lực nước lỗ rỗng và sóng âm; Cormac Reale, Kenneth Gavin và nnk [6] với đề tài nghiên cứu về phân loại tự động đất hạt mịn sử dụng kết quả xuyên tĩnh và mạng lưới thần kinh nhân tạo ANN (Artificial Neural Networks)...

Nhìn chung, việc nghiên cứu ứng dụng học máy vào công tác phân loại đất trên thế giới đang là xu hướng. Phần lớn các nghiên cứu điều tập trung khai thác bộ dữ liệu thí nghiệm xuyên tĩnh, dữ liệu địa chấn cùng với đó là ứng dụng các thuật toán bán phân cụm hoặc mạng lưới thần kinh nhân tạo ANN. Ở Việt Nam bộ dữ liệu CPT hay địa chấn chưa nhiều nhưng bộ dữ liệu về các chỉ tiêu cơ lý trong phòng của đất thì lại rất nhiều. Mục tiêu của bài báo là nghiên cứu ứng dụng thuật toán phân cụm trong phân chia đơn nguyên ĐCCT sử dụng bộ dữ liệu từ thí nghiệm trong phòng các chỉ tiêu cơ lý của đất.

Dữ liệu sử dụng trong nghiên cứu này được lấy từ dữ liệu thí nghiệm trong phòng của 437 mẫu đất ở khu vực Quận 1, Quận 8, thành phố Hồ Chí Minh phân bố từ mặt đất đến độ sâu 94.5m. Trong 437 mẫu đất

*[1]Cựu sinh viên Bộ môn Địa Kỹ thuật, Khoa Kỹ thuật Địa chất và Dầu khí, Trường Đại học Bách Khoa Tp. HCM, 268 Lý Thường Kiệt, Quận 10, Tp. Hồ Chí Minh, Việt Nam*

*[2]Bộ môn Địa kỹ thuật, Khoa Kỹ thuật Địa chất và Dầu khí, Trường Đại học Bách Khoa Tp. HCM, 268 Lý Thường Kiệt, Quận 10, Tp. Hồ Chí Minh, Việt Nam*

*[3]Đại học Quốc gia Tp. Hồ Chí Minh, Phường Linh Trung, Tp. Thủ Đức, Tp. Hồ Chí Minh, Việt Nam*

*[4]Công ty TNHH Tư vấn Địa Chất Phẳng, 85 Sương Nguyệt Anh, Quận 1, Tp. Hồ Chí Minh, Việt Nam*

**Liên hệ**

**Ngô Tấn Phong**, Bộ môn Địa kỹ thuật, Khoa Kỹ thuật Địa chất và Dầu khí, Trường Đại học Bách Khoa Tp. HCM, 268 Lý Thường Kiệt, Quận 10, Tp. Hồ Chí Minh, Việt Nam

Đại học Quốc gia Tp. Hồ Chí Minh, Phường Linh Trung, Tp. Thủ Đức, Tp. Hồ Chí Minh, Việt Nam

Email: ngotanphong@hcmut.edu.vn

**Trích dẫn bài báo này:** Thịnh H V, Chung K L T, Sơn L M, Phong N T. **Ứng dụng học máy trong phân chia đơn nguyên địa chất công trình dựa trên dữ liệu thí nghiệm đất trong phòng**. *Sci. Tech. Dev. J. - Eng. Tech.* 2024; ():1-10.

**Lịch sử**
- Ngày nhận: 01-10-2023
- Ngày chấp nhận: 25-4-2024
- Ngày đăng:

**DOI :**



này được tạo thành từ 3 bộ dữ liệu; trong đó với 2 bộ dữ liệu được thí nghiệm theo tiêu chuẩn ASTM[7] (bộ ASTM1 gồm 96 mẫu, bộ ASTM2 gồm 237 mẫu), 1 bộ dữ liệu được thí nghiệm theo tiêu chuẩn TCVN[8] gồm 104 mẫu.

# PHƯƠNG PHÁP

## Tiêu chuẩn phân loại đất

**Tiêu chuẩn ASTM D2487**

Theo tiêu chuẩn ASTM D2487, căn cứ vào đường kính cỡ hạt, ranh giới giữa đất hạt thô và đất hạt mịn là 0.075mm, đó là cỡ hạt có thể thấy được bằng mắt thường. Đầu tiên để phân loại theo ASTM này, xét phần trăm hạt qua rây 0.075mm lớn hơn hay nhỏ hơn 50%. Nếu giá trị này lớn hơn 50% thì đây là đất hạt mịn. Tiếp theo xét đến giới trị giới hạn chảy (LL), xét xem tính dẻo của loại đất lớn hơn hay nhỏ hơn 50%. Sau cùng xét chỉ số dẻo (PI) và biểu đồ Casagrande để phân loại và gọi tên đất. Nếu giá trị phần trăm hạt qua rây 0.075mm nhỏ hơn 50% đây là đất hạt thô. Xét phần trăm hạt lọt qua rây 4.75mm để biết xem đất thuộc loại sạn (phần trăm hạt giữ lại trên rây 4.75 nhiều hơn) hay loại cát (phần trăm hạt lọt qua rây 4.75mm nhiều hơn). Tiếp theo xét phần trăm hạt lọt qua rây 0.075mm và xét đến hệ số $C_u$ và $C_c$ để phân loại và gọi tên đất.

**Tiêu chuẩn TCVN 8217**

Theo tiêu chuẩn TCVN 8217, căn cứ vào đường kính cỡ hạt, ranh giới giữa đất hạt thô và đất hạt mịn là 0.08mm. Để phân loại theo TCVN, xét phần trăm các hạt có kích thước £0.08mm lớn hơn hay nhỏ hơn 50%. Nếu giá trị phần trăm này lớn hơn 50% thì đây là đất hạt mịn. Xét tiếp giới hạn chảy, giá trị chỉ số dẻo và dùng biểu đồ Casagrande để phân loại và gọi tên đất. Nếu giá trị phần trăm qua rây nhỏ hơn 50%, đây là đất hạt thô. Xét phần trăm trọng lượng thành phần hạt thô có kích thước lớn hơn 2mm. Nếu giá trị này lớn hơn 50%, đây là đất sạn sỏi, nếu hơn 50% trọng lượng thành phần hạt thô có kích thước nhỏ hơn 2mm, đây là đất cát. Xét tiếp phần trăm hạt có kích thước nhỏ hơn 0.08mm và 2 chỉ số $C_u$ và $C_c$ để phân loại và gọi tên đất.

Nhìn chung cả 2 tiêu chuẩn phân loại vừa nêu đều dựa trên thành phần hạt, tính dẻo, biểu đồ Casagrande để phân loại và gọi tên đất. Tuy nhiên, giữa 2 tiêu chuẩn có những khác biệt như trình bày ở Bảng 1.

## Thuật toán phân cụm trong học máy

K-Means, Gaussian Mixture Model (GMM), Mean Shift là các thuật toán phân cụm phổ biến trong học máy, đặc điểm của chúng được trình bày trong Bảng 2.

**Thuật toán K-Means**

Thuật toán được xây dựng dựa trên công thức:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}||x_i^{(j)} - c_j||^2 \quad (1)$$

Trong đó: k là số cụm tối ưu; n số trường hợp; $||x_i^{(j)} - c_j||^2$ là hàm khoảng cách từ 1 điểm đến điểm trung tâm. Số cụm tối ưu được xác định dựa theo biểu đồ Elbow[9].

Hàm mất mát

$$Y, M = \underset{Y,M}{argmin}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}||x_i - m_j||_2^2 \quad (2)$$

Trong đó $Y = \{y_1; y_2; ...; y_N\}$, $M = [m_1, m_2, ..., m_K]$ lần lượt là các ma trận được tạo bởi label vector của mỗi điểm dữ liệu và center của mỗi cluster.

Thỏa mãn điều kiện: $y_{ij} \in \{0,1\}\forall i, j; \sum_{j=1}^{K} y_{ij} = 1\forall i$

Thuật toán K-Means sẽ dừng lại sau một số hữu hạn vòng lặp. Hàm mất mát là một số dương và sau mỗi lần thuật toán lặp lại, giá trị hàm mất mát bị giảm đi.

**Thuật toán GMM**

Thuật toán này được xây dựng dựa trên công thức của phân phối chuẩn:

$$p(x) = \sum_{i=1}^{K}\phi_i N(x|\mu_i, \sigma_i) \quad (3)$$

$$N(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

Trong đó: $\mu_i$ là giá trị trung bình của phần tử thứ i, $\sigma_i$ là phương sai của phần tử i. Giá trị số cụm tối ưu được xác định theo biểu đồ AIC và BIC[9].

Hàm mất mát

Sử dụng thuật toán EM (Expectation-Maximization) để ước lượng hàm mất mát.

Trong thuật toán EM liên tục thực hiện các vòng lặp mà mỗi vòng lặp bao gồm hai bước huấn luyện chính:

E-Step:

Ước lượng phân phối của biến ẩn z thể hiện phân phối xác suất của các cụm tương ứng với dữ liệu và bộ tham số phân phối.

Mục tiêu của bước E-Step là tính xác suất của mỗi điểm dữ liệu dựa trên phân phối Gaussian dựa trên tham số $\theta_t$ của vòng lặp gần nhất

$$E_z(z_j|x_i, \theta_t) = \frac{\pi_i N(\mu_{jt}, \sum_{jt}|x_i)}{\sum_j \pi_j N(\mu_{jt}, \sum_{jt}|x_i)} \quad (5)$$

Trong đó: xác xuất $\pi_j$ chính là xác suất tiên nghiệm (posteriori probability) bằng với tỷ lệ các quan sát thuộc về cụm j ở vòng lặp thứ t. $N(\mu_{jt}, \sum_{jt}|x_i)$ là xác suất của $x_i$ rơi vào cụm thứ j được tính theo phân phối Gaussian đa chiều.

M-Step:

Tối đa hóa phân phối xác suất đồng thời (join distribution probability) của dữ liệu và biến ẩn. Tại bước

**Bảng 1: Điểm khác nhau giữa tiêu chuẩn ASTM D2487 và TCVN 8217**

| Tiêu chuẩn | ASTM D2487 | TCVN 8217 |
|---|---|---|
| Ranh giới cỡ hạt mịn và thô, mm | 0.075 | 0.080 |
| Mẫu đất xác định giới hạn chảy phải lọt qua rây, mm | 0.425 | 1.000 |
| Khối lượng chùy xuyên trong thí nghiệm xác định giới hạn chảy, g | 80 | 76 |
| Thời gian côn xuyên vào đất trong thí nghiệm xác định giới hạn chảy, giây | 5 | 10 |

**Bảng 2: Đặc điểm của 3 thuật toán K-Means, GMM và Mean Shift [9]**

| K-Means | Gaussian Mixture Model | Mean Shift |
|---|---|---|
| Dựa trên điểm trung tâm | Dựa trên sự phân bố dữ liệu | Dựa trên mật độ dữ liệu |
| Bị ảnh hưởng nhiều bởi giá trị ngoại lai | Ít bị ảnh hưởng bởi giá trị ngoại lai | Ít bị ảnh hưởng bởi giá trị ngoại lai |
| Cần tìm số cụm tối ưu (sử dụng phương pháp Elbow) | Cần tìm số cụm tối ưu (sử dụng 2 chỉ số AIC và BIC) | Không cần tìm số cụm tối ưu |

M-Step cập nhật lại tham số phân phối theo hàm auxiliary $Q(\theta, \theta_t)$

$$\pi_j^* = \frac{\sum_{i=1}^{N} p\left(z_j | x_i, \theta_t\right)}{N} \quad (6)$$

**Thuật toán Mean Shift**

Thuật toán Mean Shift được xây dựng trên công thức hàm mật độ:

$$\widehat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (7)$$

Trong đó n: là số điểm dữ liệu; d: kích thước dữ liệu; h là giá trị bandwith [9].

Hàm mất mát

$$\nabla_x \widehat{f}(x) = \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g(y)\right] \left[\frac{x_i}{\sum_{i=1}^{n} g(y)} - x\right] \quad (8)$$

Trong đó: $\frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g(y)\right]$ là tỷ trọng phân phối của hàm mật độ xác suất tại điểm x.

$\left[\frac{x_i}{\sum_{i=1}^{n} g(y)} - x\right]$ được gọi là meanshift vector

Để đơn giản hóa việc tính toán, toàn bộ 3 thuật toán phân cụm được tác giả sử dụng theo bộ thư viện sklearn thuộc ngôn ngữ lập trình Python [10].

## Quy trình phân chia đơn nguyên ĐCCT

Quy trình phân chia đơn nguyên ĐCCT được thực hiện theo 5 bước sau: **Bước 1.** Lựa chọn các thông số đầu vào mô hình; **Bước 2.** Chuẩn hóa dữ liệu (ứng dụng hàm standardScaler() thuộc bộ thư viện sklearn trong ngôn ngữ lập trình Python); **Bước 3.** Tìm số cụm tối ưu; **Bước 4.** Chạy mô hình thuật toán phân cụm; **Bước 5.** Xuất kết quả phân lớp.

Các thông số được sử dụng trong phân cụm gồm thành phần hạt, giới hạn chảy (LL), giới hạn dẻo (PL), độ ẩm (MC), dung trọng tự nhiên (DEN), hệ số rỗng (VOID), các thông số được trình bày chi tiết ở Bảng 3. Lưu ý về thành phần hạt, các thông số liên quan đến hàm lượng kích thước hạt cuội, sạn, cát, bụi và sét được xem như là mỗi thông số riêng biệt trong quá trình phân cụm. Có tổng cộng 17 thông số tương ứng với các bộ dữ liệu ASTM1, ASTM2 và TCVN.

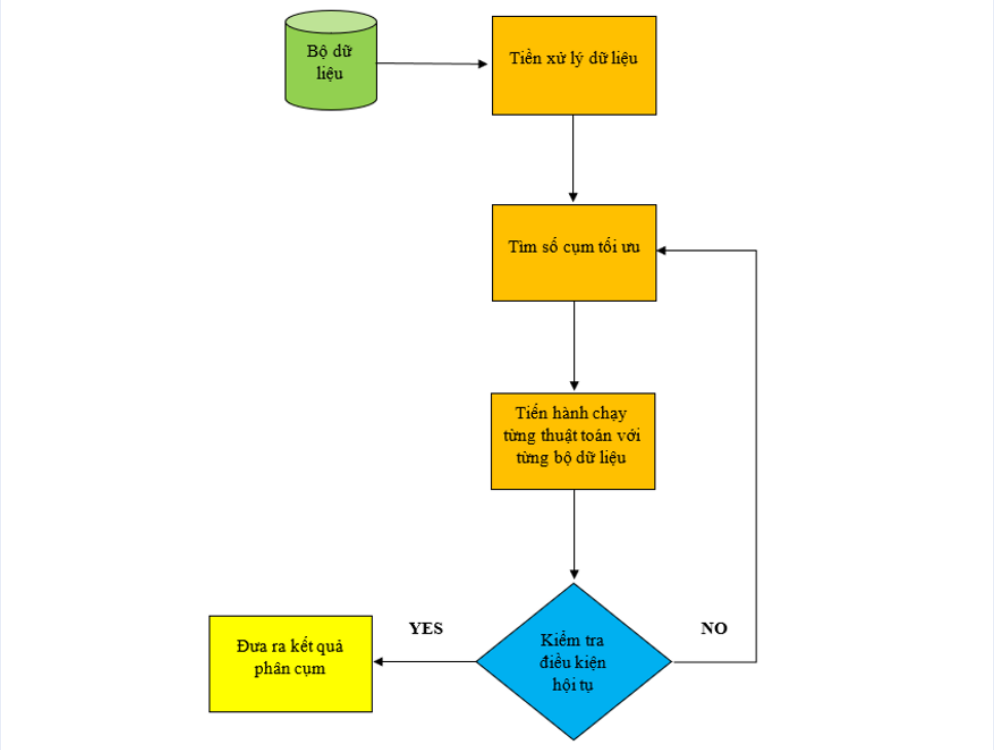
Để đánh giá mức độ ảnh hưởng của thông số đầu vào đến độ chính xác của mô hình phân cụm, bài báo đã xem xét thay đổi số lượng thông số đầu vào lần lượt là 12, 14 và 17 tương ứng với ba trường hợp TH1, TH2 và TH3. Trường hợp TH1, có 12 thông số đầu vào được xem xét liên quan đến thành phần hạt. Trường hợp TH2, có 14 thông số đầu vào được xem xét liên quan đến thành phần hạt, giới hạn chảy và giới hạn dẻo. Trường hợp TH3, có 17 thông số đầu vào được xem xét liên quan đến thành phần hạt, giới hạn chảy, giới hạn dẻo, độ ẩm, dung trọng tự nhiên và hệ số rỗng.

**Chú thích:** TH1-thông số đầu vào chỉ gồm thành phần hạt (GR1, GR2, SA6, SA5, SA4, SA3, SA2, SA1, SI2, SI1, CL2, CL1) trong trường hợp bộ dữ liệu ASTM1. Tương tự cho các bộ dữ liệu khác. TH2-thông số đầu vào gồm thành phần hạt, giới hạn chảy, giới hạn dẻo (TH1+LL+PL). TH3-thông số đầu vào gồm thành phần hạt, giới hạn chảy, giới hạn dẻo, độ ẩm, dung trọng tự nhiên, hệ số rỗng (TH2+MC+DEN+VOID)

Sơ đồ tính toán được trình bày trong Hình 1.

**Bảng 3: Các thông số sử dụng khi phân cụm với bộ dữ liệu ASTM và TCVN**

| ASTM1 | | ASTM2 | | TCVN | |
|---|---|---|---|---|---|
| GR1-sạn | SI1-bụi | GR1-sạn | SI1-bụi | CO1-cuội | SI2-bụi |
| GR2-sạn | CL2-sét | GR2-sạn | CL2-sét | GR3-sạn | SI1-bụi |
| SA6-cát vừa | CL1-sét | SA6-cát thô | CL1-sét | GR2-sạn | CL1-sét |
| SA5-cát vừa | LL-giới hạn chảy | SA5-cát vừa | LL-giới hạn chảy | GR1-sạn | LL-giới hạn chảy |
| SA4-cát vừa | PL-giới hạn dẻo | SA4-cát vừa | PL-giới hạn dẻo | SA5-cát thô | PL-giới hạn dẻo |
| SA3-cát mịn | MC (w)-độ ẩm | SA3-cát mịn | MC-độ ẩm | SA4-cát vừa | MC-độ ẩm |
| SA2-cát mịn | DEN-dung trọng tự nhiên | SA2-cát mịn | DEN-dung trọng tự nhiên | SA3-cát mịn | DEN-dung trọng tự nhiên |
| SA1-cát mịn | VOID-hệ số rỗng | SA1-cát mịn | VOID-hệ số rỗng | SA2-cát mịn | VOID-hệ số rỗng |
| SI2-bụi | | SI2-bụi | | SA2-cát mịn | |




**Hình 1**: Sơ đồ tính toán

# KẾT QUẢ VÀ THẢO LUẬN

## Độ chính xác

Độ chính xác được tính toán dựa trên việc so sánh kết quả phân lớp theo phương pháp thủ công và phân lớp theo thuật toán phân cụm. Độ chính xác được xác định từ số mẫu phân lớp giống nhau giữa 2 phương pháp (a) và tổng số mẫu (b) như sau:

Độ chính xác = $\frac{a}{b} \times 100\ (\%)$ (9)

Trong đó: a là số mẫu phân lớp giống nhau giữa phương pháp học máy và thủ công; b là tổng số mẫu.

Độ chính xác của 3 thuật toán K-Means, GMM, Mean Shift tương ứng với 3 bộ dữ liệu ASTM1, ASTM2, TCVN trong trường hợp TH3 được biểu thị ở hình Hình 2. Lưu ý, trường hợp TH3 (17 thông số) cho độ chính xác cao nhất. Dựa vào đồ thị này có thể thấy, theo xu hướng chung, giá trị độ chính xác tăng theo trình tự từ thuật toán K-Means đến Mean Shift và độ chính xác đạt giá trị cao nhất là 93.8% ở thuật toán GMM. Độ chính xác của thuật toán GMM cao gấp 1.07 lần Mean Shift và gấp 1.05 lần K-Means.

Từ kết quả thuật toán GMM phân tích kết hợp với bảng phân loại thủ công tiến hành vẽ các hình trụ hố khoan.

Quan sát hình Hình 3a dễ nhận thấy kết quả phân chia theo phương pháp phân cụm khá tương đồng với phân chia theo phương pháp thủ công. Tuy nhiên một số mẫu ở độ sâu từ 15-18.5m kết quả thuật toán phân cụm có sự khác biệt. Nguyên nhân của sự khác biệt này là do tại vị trí đó có khá ít điểm dữ liệu. Ở Hình 3b và Hình 3c cho kết quả tương đồng giữa phân chia theo học máy và phân chia thủ công.

## Ảnh hưởng số lượng thông số đầu vào

Nhằm đánh giá ảnh thưởng của số lượng thông số đầu vào đến độ chính xác của mô hình phân cụm, số lượng thông số đầu vào được thay đổi lần lượt là 12, 14 và 17 tương ứng với ba trường hợp TH1, TH2 và TH3. Kết quả độ chính xác của 3 thuật toán K-Means, GMM và Mean Shift qua 3 bộ dữ liệu tương ứng với các trường hợp thông số mô hình được thể hiện trong các Bảng 4, 5 và 6.

Độ chính xác của thuật toán K-Means tăng theo số lượng thông số đầu vào mô hình. Dựa vào đồ thị trên Hình 4 có thể thấy, theo xu hướng chung khi số lượng thông số mô hình càng tăng thì độ chính xác của thuật toán cũng tăng theo. Độ chính xác đạt giá trị lớn nhất tại trường hợp TH3 (17 thông số) là 88.5%. Độ chính xác tăng từ 83.3% lên 88.5% khi số lượng thông số tăng từ 12 lên 17 thông số.

Độ chính xác của thuật toán GMM qua 3 trường hợp thông số mô hình được thể hiện trong Hình 5. Dựa vào đồ thị dễ nhận thấy, độ chính xác của thuật toán tăng khi số lượng thông số mô hình tăng lên. Cụ thể độ chính xác tăng từ 82.3% lên 93.8% khi số lượng thông số tăng từ 12 lên 17 thông số.

Đồ thị trên Hình 6 thể hiện độ chính xác của thuật toán Mean Shift qua 3 trường hợp TH1, TH2 và TH3. Tương tự như ở đồ thị Hình 4 và Hình 5, dễ dàng nhận thấy 1 xu hướng chung, trong trường hợp này độ chính xác tăng theo số lượng thông số mô hình. Tuy nhiên, xuất hiện dị thường với trường hợp TH2 và TH3. Trường hợp TH2 có một vài điểm bằng thậm chí cao hơn trường hợp TH3. Nguyên nhân có thể là do bộ dữ liệu dùng trong đề tài chưa đủ lớn cùng với đó sự chênh lệch độ chính xác giữa 2 trường hợp là không cao bởi lẽ mỗi lần thuật toán chạy lại thì độ chính xác cũng có sự thay đổi nhỏ.

Về xu hướng chung, có thể kết luận độ chính xác của 3 thuật toán phân cụm có mối quan hệ tương quan thuận với số lượng thông số đầu vào mô hình. Khi số lượng thông số mô hình tăng thì độ chính xác của các thuật toán sẽ tăng theo.

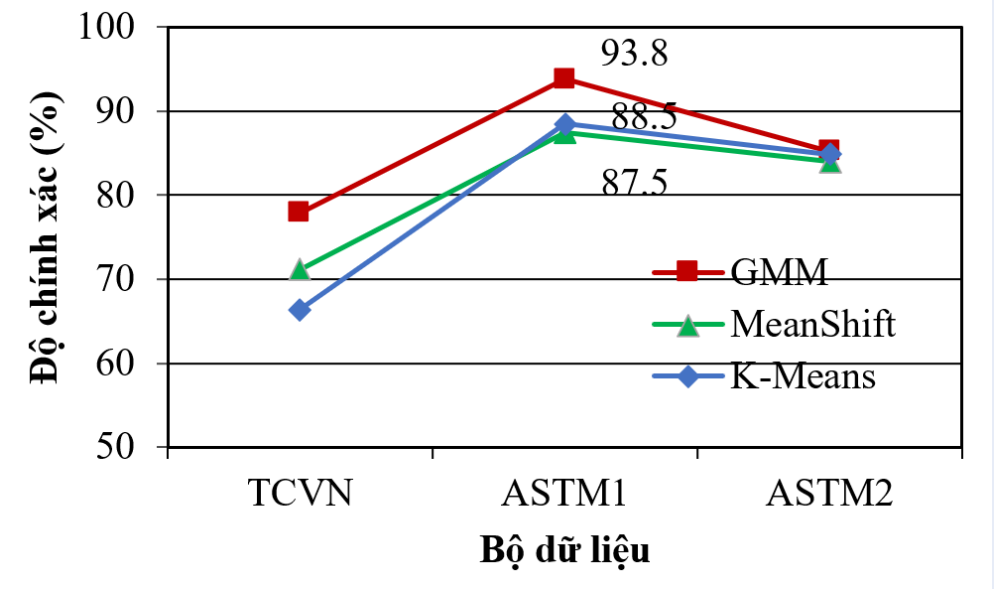
## Dị thường

Dựa vào Bảng 7 có thể dễ dàng nhận thấy hàm lượng hạt bụi trong 2 mẫu 49 và 95 cao hơn các mẫu còn lại. Do đó thuật toán đã phân tách các mẫu này vào một lớp riêng biệt. Ở phân chia theo phương pháp thủ công với các đặc tính mẫu đất như vậy, tất cả mẫu này chỉ được phân vào cùng 1 lớp duy nhất, lớp ký hiệu MH. Tuy nhiên khi sử dụng thuật toán GMM, thuật toán đã phân chia lớp này ra thành 2 lớp nhỏ hơn. Như vậy thuật toán phân cụm có thể giúp phân chia chi tiết hơn các lớp đất, từ đó giúp phát hiện các lớp kẹp mỏng trong địa tầng.

Tương tự, ở các mẫu 54 và 100 trong Bảng 8 có hàm lượng cát từ hạt thô đến hạt vừa cao hơn ở các mẫu còn lại**.** Nếu ở phân chia theo phương pháp thủ công, tất cả các mẫu này sẽ được gom vào cùng một lớp là cát bụi, SM. Tuy nhiên khi phân chia theo thuật toán GMM thì các mẫu này được chia ra thành 2 lớp, điều mà phương pháp thủ công chưa làm được.

Nhìn chung độ chính xác của các thuật toán phân cụm sẽ tăng theo số lượng thông số mô hình nhưng theo kết quả từ hình Hình 6 có thể thấy ở bộ dữ liệu ASTM1 xuất hiện dị thường. Trường hợp TH3 có số lượng thông số mô hình cao hơn trường hợp TH2 nhưng lại có độ chính xác thấp hơn trường hợp TH2. Nguyên nhân có thể là do số lượng mẫu dùng trong nghiên cứu khá ít dẫn đến thuật toán chưa nhận diện được chính xác.

Lưu ý độ chính xác của thuật toán sẽ thay đổi ở mỗi lần chạy vì thuật toán giả định điểm trung tâm ở mỗi lần chạy tương ứng. Do vậy giá trị độ chính xác sẽ thay đổi, tuy nhiên giá trị thay đổi này là rất nhỏ, không đáng kể có thể bỏ qua khi phân tích.




**Hình 2**: So sánh độ chính xác giữa 3 thuật toán phân cụm đối với bộ dữ liệu ASTM1, ASTM2 và TCVN trong trường hợp TH3

**Bảng 4: Độ chính xác của 3 thuật toán trường hợp TH1 (12 thông số)**

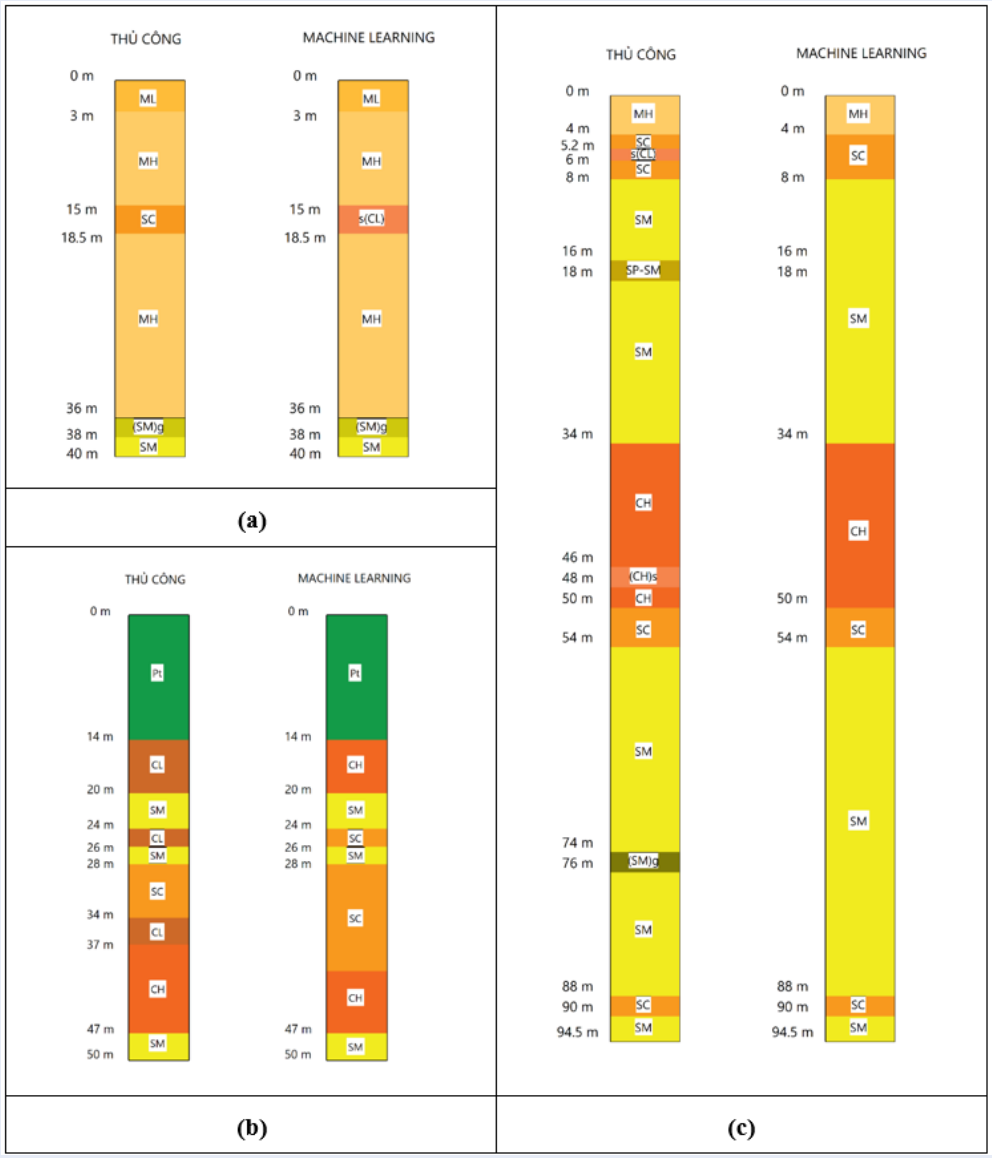
| Thuật toán | K-Means | GMM | Mean Shift |
|---|---|---|---|
| Bộ dữ liệu | | | |
| TCVN | 59.6% | 60.6% | 56.7% |
| ASTM1 | 83.3% | 82.3% | 85.4% |
| ASTM2 | 77.2% | 78.1% | 79.8% |

**Bảng 5: Độ chính xác của 3 thuật toán trường hợp TH2 (14 thông số)**

| Thuật toán | K-Means | GMM | Mean Shift |
|---|---|---|---|
| Bộ dữ liệu | | | |
| TCVN | 61.5% | 68.3% | 66.4% |
| ASTM1 | 84.4% | 91.7% | 90.6% |
| ASTM2 | 78.1% | 81.0% | 84.0% |

**Bảng 6: Độ chính xác của 3 thuật toán trường hợp TH3 (17 thông số)**

| Thuật toán | K-Means | GMM | Mean Shift |
|---|---|---|---|
| Bộ dữ liệu | | | |
| TCVN | 66.4% | 77.9% | 71.2% |
| ASTM1 | 88.5% | 93.8% | 87.5% |
| ASTM2 | 84.8% | 85.2% | 84.0% |

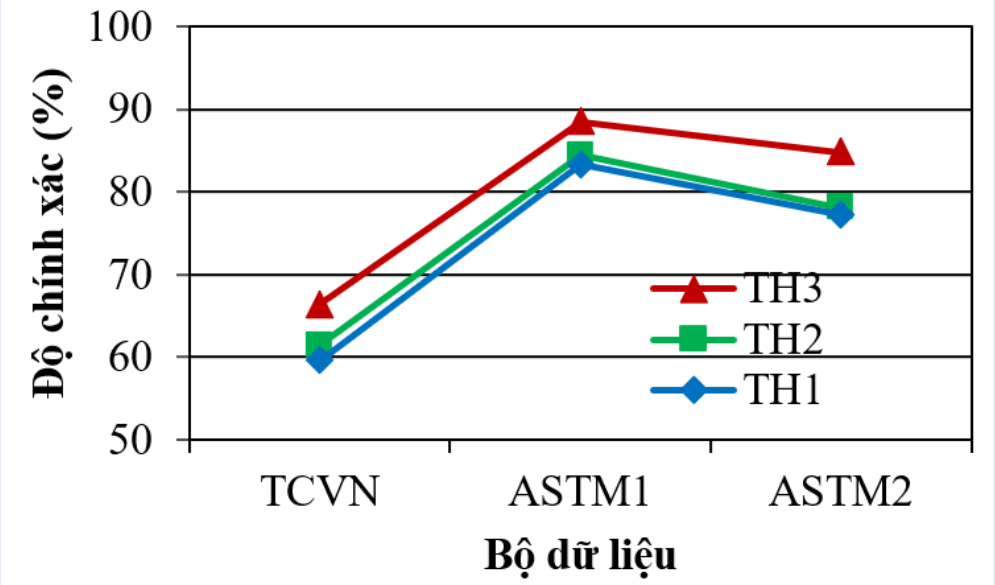



**Hình 3**: (a) Hình trụ hố khoan đối với bộ dữ liệu ASTM1 trường hợp TH3; (b) Hình trụ hố khoan đối với bộ dữ liệu TCVN trường hợp TH3; (c) Hình trụ hố khoan đối với bộ dữ liệu ASTM2 trường hợp TH3

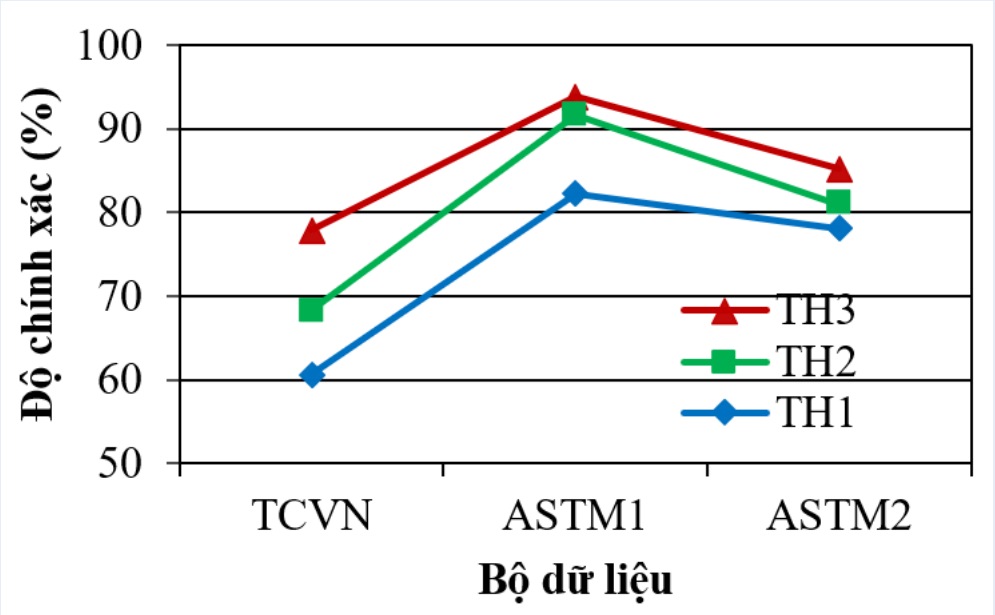
## KẾT LUẬN

Trên cơ sở ứng dụng thuật toán phân cụm K-Means, GMM và Mean Shift trong việc phân chia đơn nguyên ĐCCT cho bộ dữ liệu các chỉ tiêu cơ lý của đất, bài báo đi đến các kết luận sau:

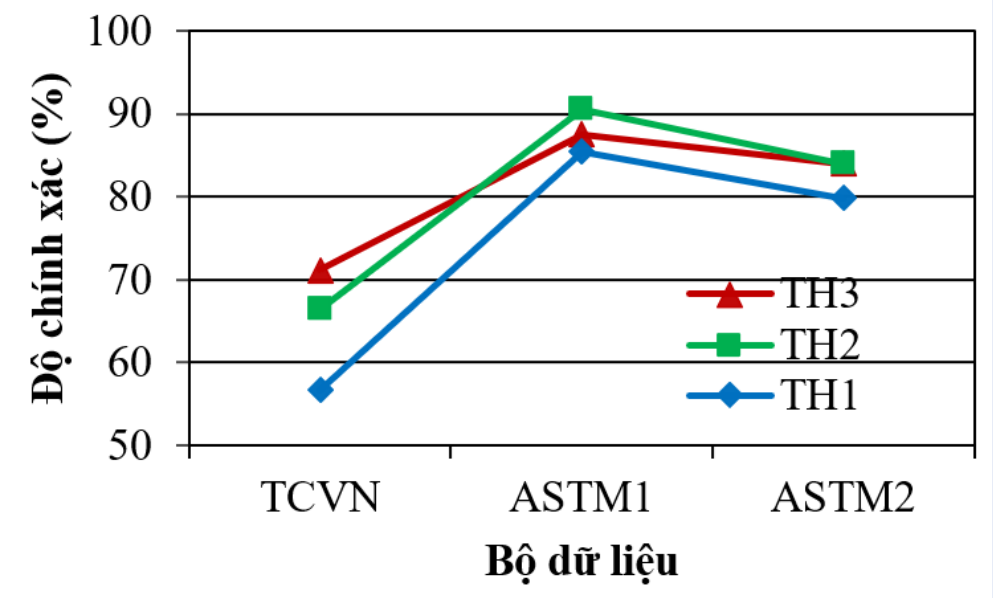
- Trong 3 thuật toán phân cụm K-Means, GMM và Mean Shift thì thuật toán GMM cho kết quả với độ chính xác cao nhất là 93.8% tương ứng với bộ dữ liệu ASTM1 trường hợp TH3 (17 thông số).
- Nhìn chung bộ dữ liệu ASTM cho kết quả tốt hơn bộ dữ liệu TCVN. Trong đó bộ dữ liệu ASTM đạt độ chính xác cao nhất là 93.8% trong khi bộ TCVN chỉ đạt 77.9%. Nguyên nhân có thể là do sự khác nhau về ranh giới cỡ hạt, phương pháp thí nghiệm xác định giới hạn chảy. Do vậy có thể kết luận rằng bộ dữ liệu ASTM phù hợp hơn cho thuật toán GMM.
- Thuật toán phân cụm có thể xác định được các dị thường bằng cách so sánh các tính chất của một mẫu đất này với các tính chất của mẫu đất khác. Thuật toán có thể phân chia các lớp đất chi




**Hình 4**: Độ chính xác của thuật toán K-Means đối với thông số mô hình trong 3 trường hợp TH1, TH2, TH3




**Hình 5**: Độ chính xác của thuật toán GMM đối với thông số mô hình trong 3 trường hợp TH1, TH2, TH3




**Hình 6**: Độ chính xác của thuật toán Mean Shift đối với thông số mô hình trong 3 trường hợp TH1, TH2, TH3

**Bảng 7: Kết quả phân chia đơn nguyên sử dụng thuật toán GMM ở lớp bụi tính dẻo cao (MH) của bộ dữ liệu ASTM2 trường hợp TH3**

| Số hiệu mẫu | SI2-bụi (%) | SI1-bụi (%) | Hàm lượng hạt bụi (%) | Thủ công | Machine learning |
|---|---|---|---|---|---|
| - | 0.075 ÷ 0.02 mm | 0.02 ÷ 0.005 mm | - | - | - |
| 0 | 16.1 | 22 | 38.1 | MH | 7 |
| 1 | 10.3 | 23.4 | 33.7 | MH | 7 |
| 48 | 16.2 | 19.7 | 35.9 | MH | 7 |
| 49 | 28.1 | 22.2 | 50.3 | MH | 3 |
| 95 | 32.5 | 24.2 | 56.7 | MH | 3 |
| 142 | 16.9 | 20.7 | 37.6 | MH | 7 |

**Ghi chú:** giá trị i trong cột Machine learning nghĩa là lớp đất i.

tiết hơn từ việc phát hiện ra các dị thường trong bộ dữ liệu, điều mà phương pháp phân chia thủ công chưa làm được.

- Số lượng thông số mô hình có mối quan hệ tương quan thuận với độ chính xác của thuật toán. Khi số lượng thông số mô hình tăng thì độ chính xác của thuật toán sẽ tăng theo. Độ chính xác của thuật toán GMM ở bộ dữ liệu ASTM1 tăng từ 82.3% lên 93.8% khi số lượng thông số tăng từ 12 lên 17 thông số.

Nghiên cứu này chỉ tiến hành phân cụm trên dữ liệu về các chỉ tiêu cơ lý (thành phần hạt, giới hạn chảy, giới hạn dẻo, độ ẩm, dung trọng tự nhiên, hệ số rỗng) của đất, chưa sử dụng dữ liệu về màu sắc, trạng thái và độ sâu của lớp đất. Do đó cần bổ sung thêm các nghiên cứu sử dụng các thông số về màu sắc, trạng thái và độ sâu của lớp đất trong thuật toán phân cụm để phân chia đơn nguyên ĐCCT trong tương lai. Bên cạnh đó, nghiên cứu này chỉ dựa trên dữ liệu các mẫu đất ở Quận 1 và Quận 8 thuộc Tp. Hồ Chí Minh. Vì vậy, trong tương lai cần mở rộng phạm vi nghiên cứu ở các khu vực khác để đánh giá khách quan về mức độ hiệu quả của phương pháp Machine learning trong phân loại đất.

**Bảng 8: Kết quả phân chia đơn nguyên sử dụng thuật toán GMM ở lớp cát bụi (SM) của bộ dữ liệu ASTM2 trường hợp TH3**

| Số hiệu mẫu | SA6-cát thô (%) | SA5-cát vừa (%) | SA4-cát vừa (%) | Hàm lượng cát (thô-vừa) (%) | Thủ công | Machine learning |
|---|---|---|---|---|---|---|
| - | 4.75 ÷ 2.00 mm | 2.00 ÷ 0.85 mm | 0.85 ÷ 0.425 mm | - | - | - |
| 5 | 0 | 0 | 4.8 | 4.8 | SM | 2 |
| 99 | 0.2 | 1.1 | 5.2 | 6.5 | SM | 2 |
| 147 | 0.3 | 2.3 | 12.3 | 14.9 | SM | 2 |
| 54 | 3.6 | 16.4 | 18.8 | 38.8 | SM | 4 |
| 6 | 0.3 | 0.3 | 9.7 | 10.3 | SM | 2 |
| 53 | 0 | 0 | 3.9 | 3.9 | SM | 2 |
| 100 | 2.3 | 22.8 | 21.7 | 46.8 | SM | 4 |

**Ghi chú:** giá trị i trong cột Machine learning nghĩa là lớp đất i.

## XUNG ĐỘT LỢI ÍCH

Nhóm tác giả cam kết không mâu thuẫn về quyền lợi và nghĩa vụ của các thành viên.

## ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Tác giả Huỳnh Văn Thịnh thực hiện nhiệm vụ lập trình python, phân tích kết quả và viết bản nháp bài báo; tác giả Kiều Lê Thủy Chung tổng hợp dữ liệu thí nghiệm đất, hỗ trợ hiệu chỉnh nội dung; tác giả Lê Minh Sơn đề xuất ý tưởng, cung cấp dữ liệu, chỉnh sửa code python và góp ý chỉnh sửa; tác giả Ngô Tấn Phong thực hiện chỉnh sửa toàn văn bài báo, liên hệ, nộp và trả lời câu hỏi của phản biện và ban biên tập.

## TÀI LIỆU THAM KHẢO


1. Bình HV, Hảo LP. Hệ thống phân loại đất đá phục vụ cho xây dựng các công trình thủy điện khu vực Tây Nguyên. Tạp chí phát triển Khoa học và Công nghệ. 2007;10;.
2. Dư TT, Ninh OV, Vũ PT, et al. Các đặc tính phân loại đất theo hệ thống phân loại WRB 2006 tỉnh Hậu Giang tỉ lệ 1/100.000. Đại học Cần Thơ; 2006;.
3. Minh VQ, Trí LQ. Đất Đồng bằng sông Cửu Long phân loại theo hệ thống WRB. Đại học Cần Thơ; 2006;.
4. Wang X, Wang H, Liang RY, Liu Y. A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data. Eng Geol. 2019;248:102-116;Available from: https://doi.org/10.1016/j.enggeo.2018.11.014.
5. Molina-Gómez F, Viana da Fonseca A, et al. Defining the soil stratigraphy from seismic piezoncone data: A clustering approach. Eng Geol. 2021;287:106-111;Available from: https://doi.org/10.1016/j.enggeo.2021.106111.
6. Reale C, Gavin K, et al. Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks. Adv Eng Informatics. 2018;36:207-215;Available from: https://doi.org/10.1016/j.aei.2018.04.003.
7. ASTM International. Standard practice for classification of soils for engineering purposes (Unified Soil Classification System) (ASTM D2487-17). West Conshohocken, PA, USA: ASTM International; 2017;.
8. Tổng cục Tiêu chuẩn Đo lường Chất lượng. TCVN 5747:1993. Đất xây dựng - Phân loại. Hà Nội: Bộ Khoa học Công nghệ; 1993;.
9. Vanderplas J. Python data science handbook. University of Washington; 2016;.
10. Tiệp VH. Machine learning cơ bản. Hà Nội: Nhà xuất bản Khoa học và Kỹ thuật; 2018;.

*Science & Technology Development Journal – Engineering and Technology 2024, ():1-1*

Open Access Full Text Article

*Research Article*

# Applying machine learning for soil classification in geotechnical engineering based on laboratory soil test data

**Huynh Van Thinh[1], Kieu Le Thuy Chung[2,3], Le Minh Son[4], Ngo Tan Phong[2,3,*]**



**ABSTRACT**

An engineering geological unit/soil layer is a homogeneous volume of soil and rock with the same name and physical and mechanical characteristics that vary without regularity, or if the physical and mechanical characteristics vary with a regularity then this regularity can be ignored when statistical conditions are satisfied. Classificating soil layers is very important to provide reliable data for calculating, designing the foundations, and choosing reasonable construction solutions that contribute to saving the costs of building constructions. Currently, the classification of soil layers is often carried out manually based on soil classification standards of Vietnam (TCVN), America (ASTM), Britain (BS), etc. Because the work is carried out manually, errors are inevitable, especially when working with big data sources. To reduce errors when classification of soil layers as well as detect anomalies in the soil stratigraphy and save time when synthesizing large amounts of data to build a geotechnical database for a region, this article aims to apply Machine learning to automatically classify soil layers based on the use of three clustering algorithms such as K-Means, Gaussian Mixture Model (GMM), and Mean Shift. The input data set is taken from the results of soil testing of 437 soil samples taken from District 1 and District 8 in Ho Chi Minh City. The results show that the automatic classification program gives results of soil layers that closely match the results of the manual method. Furthermore, the automatic program can divide soil stratigraphy into very detailed units, which helps detect anomalies in geotechnical engineering.

**Key words:** Gaussian Mixture Models, K-Means, Machine learning, Mean Shift, Soil classification

[1]*Former student, Department of Geotechnics, Faculty of Geology & Petroleum Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam*

[2]*Department of Geotechnics, Faculty of Geology & Petroleum Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam*

[3]*Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam*

[4]*flatGEO Consulting Company, 85 Suong Nguyet Anh Street, District 1, Ho Chi Minh City, Vietnam*

**Correspondence**

**Ngo Tan Phong**, Department of Geotechnics, Faculty of Geology & Petroleum Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

Email: ngotanphong@hcmut.edu.vn

**Cite this article :** Thinh H V, Chung K L T, Son L M, Phong N T. **Applying machine learning for soil classification in geotechnical engineering based on laboratory soil test data**. *Sci. Tech. Dev. J. – Engineering and Technology* 2024; ():1-1.