Tạp chí Khoa học Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.343

PHÂN LOẠI CHO CÁC HÀM MẬT ĐỘ XÁC SUẤT VÀ ỨNG DỤNG CHO ẢNH

Trần Nguyễn Kim Ngân[1], Võ Thị Cẩm Tiên[1], Lê Thanh Tâm[1], Nguyễn Phúc Bảo[1], Nguyễn Thị Mẫn Trâm[1], Lê Thị Huỳnh Như[1], Nguyễn Thị Yến Nhi[1], Thái Minh Trọng[1] và Lê Đại Nghiệp[2*]

[1]*Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[2]*Khoa Cơ bản, Trường Đại học Nam Cần Thơ*
**Tác giả liên hệ (Corresponding author): ldnghiep@nctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 01/05/2024
Sửa bài (Revised): 28/06/2024
Duyệt đăng (Accepted): 01/08/2024

Title: *Classifying for probability density functions and applying for images*

Author(s): *Tran Nguyen Kim Ngan[1], Vo Thi Cam Tien[1], Le Thanh Tam[1], Nguyen Phuc Bao[1], Nguyen Thi Man Tram[1], Le Thi Huynh Nhu[1], Nguyen Thi Yen Nhi[1], Thai Minh Trong[1] and Le Dai Nghiep[2*]*

Affiliation(s): *[1]Can Tho University, [2]Nam Can Tho University*

TÓM TẮT

Nghiên cứu này đề xuất một thuật toán phân loại cho các hàm mật độ xác suất (PDF) để từ đó áp dụng cho dữ liệu ảnh. Thuật toán đề nghị được trình bày chi tiết các bước thực hiện và được minh hoạ trên một tập PDF cụ thể. Để áp dụng cho dữ liệu ảnh, nghiên cứu trích xuất đặc trưng màu sắc với 4 màu cơ bản thành các PDF một chiều đại diện. Sau đó, phương pháp tìm xác suất tiên nghiệm dựa trên kỹ thuật phân tích chùm mờ được xây dựng. Cuối cùng, nguyên tắc phân loại tựa Bayes được thiết lập. Ứng dụng trên tập ảnh cụ thể cho thấy kết quả phân loại tốt và có nhiều tiềm năng trong áp dụng thực tế của nhiều lĩnh vực khác nhau.

Từ khoá: *Hàm mật độ xác suất, phân loại, phương pháp Bayes, trích xuất ảnh*

ABSTRACT

This study proposes a classification algorithm for probability density functions (PDFs) which is then applied to image data. The proposed algorithm is detailed step-by-step and illustrated using a specific set of PDFs. For application to image data, the study extracts color features using four basic colors represented as one-dimensional PDFs. Next, the method for finding prior probabilities based on fuzzy cluster analysis techniques is built. Finally, the quasi-Bayesian classification principle is established. Application on a specific set of images shows good classification results and demonstrates significant potential for practical applications across various fields.

Keywords: *Probability density function, classification, Bayesian method, extracting image*

1. GIỚI THIỆU

Theo Pham-Gia et al. (2000), nhận dạng thống kê gồm có hai bài toán chính: nhận dạng không được giám sát và nhận dạng được giám sát. Nhận dạng không được giám sát là việc chia một tập dữ liệu đã cho thành các chùm sao cho các phần tử trong cùng một chùm có sự tương tự nhiều hơn so với các phần tử khác bên ngoài chùm, căn cứ vào các biến quan sát của nó (Dinh & Tai, 2021, 2023). Do đó, nhận dạng không được giám sát còn được gọi là bài toán phân tích chùm. Nhận dạng được giám sát là việc xếp một phần tử vào một trong các nhóm đã cho một cách thích hợp nhất. Chính vì vậy, nhận dạng được giám sát được gọi là bài toán phân loại (Thao & Tai, 2020; Ha et al., 2020). Phân loại là một hướng phát triển quan trọng của thống kê nhiều chiều và khoa

học dữ liệu ngày nay. Nó đã và đang được áp dụng trong nhiều vấn đề cụ thể trong kinh tế, xã hội, y học và môi trường.

Phân loại có thể thực hiện cho dữ liệu rời rạc, khoảng và hàm mật độ xác suất (PDF). Dữ liệu rời rạc được xây dựng đầu tiên cho bài toán phân loại và được áp dụng cho nhiều vấn đề của các lĩnh vực khác nhau (Katarzyna, 2017; Tai et al., 2022;). Trong thực tế, chúng ta cũng lưu trữ nhiều dữ liệu khoảng như giá trị cao nhất và thấp nhất trong ngày của nhiệt độ, độ ẩm, chỉ số nắng,… Do đó, phân loại cho dữ liệu khoảng đã được đề xuất. Với dữ liệu này, một số thuật toán quan trọng gần đây đã được đề xuất (Ngoc et al., 2022; Dan & Tai, 2024). Khi dữ liệu lớn, một đối tượng có thể biểu diễn thành một PDF, nên phân loại cho đối tượng này đã được đề nghị. Hơn nữa, một ảnh có thể cũng được biểu diễn thành một PDF, nên việc phân loại cho đối tượng này là cần thiết và rất quan trọng.

So với dữ liệu số và khoảng, việc phân loại cho các PDF còn rất nhiều hạn chế, hầu như chưa được quan tâm bởi các nhà thống kê. Hơn nữa, việc phân loại cho đối tượng này được đánh giá rất phức tạp. Sự phức tạp được thể hiện trong tất cả các bước quan trọng chính. Đầu tiên, đó là việc xây dựng các độ đo đánh giá sự tương tự của hai phần tử và giữa một phần tử đối với các nhóm. Với dữ liệu rời rạc, nhiều khoảng cách khác nhau đã được đề xuất và áp dụng như khoảng cách Euclide, City-block, khoảng cách L^p. Đối với dữ liệu khoảng, khoảng cách Huassdoff, City-block và L^p thường được áp dụng. Cho dữ liệu PDF, có rất nhiều khoảng cách khác nhau được định nghĩa dựa trên sự mở rộng từ đối tượng rời rạc. So với dữ liệu rời rạc và khoảng, các độ đo cho PDF thường rất khó trong tính toán để áp dụng trong thực tế. Một vấn đề phức tạp khác trong phân loại PDF là việc xây dựng nguyên tắc phân loại. Nguyên tắc phân loại đối với dữ liệu rời rạc thì rất phong phú. Đó là những nguyên tắc phân loại được xây dựng dựa vào thống kê như Logistic, Naive Bayes, Fisher (Fisher, 1938; Che-Ngoc, 2023) và dựa vào học máy như kNN, máy véctơ hỗ trợ (SVM), Trees, Neural Networks, Optimal Autofit, SqueezeNet, GoogleNet, MobileNetv2, VGG-19 và Inceptionv (Imandoust et al., 2013; Pham et al., 2016; Huang et al., 2018; Koklu, 2021; Philipp et al., 2021; Cinar, 2022). Đối với dữ liệu khoảng, chúng ta có thể xây dựng nguyên tắc phân loại tương tự như dữ liệu rời rạc. Tuy nhiên sự phát triển từ dữ liệu rời rạc và khoảng cho PDF rất khó được thực hiện.

Với dữ liệu ảnh, để áp dụng bài toán phân loại, chúng ta đầu tiên phải trích xuất các đặc trưng của nó. Có nhiều phương pháp khác nhau để trích xuất đặc trưng của ảnh, trong đó dựa vào màu sắc là phương pháp phổ biến. Trong nhiều nghiên cứu, từ 3 màu cơ bản RGB (Đỏ, Xanh lục, Xanh lam), một ảnh bất kỳ sẽ được chuyển về màu Gr (Màu xám) để trích xuất đặc trưng (Zhang et al., 2018). Sau khi ma trận đặc trưng màu xám được trích xuất, các biến rời rạc và khoảng được thiết lập để đại diện cho một ảnh (Ngoc et al., 2021; Che-Ngoc et al., 2023). Việc phân biệt các ảnh với nhau chính là dựa trên các đối tương đại diện này. Việc sử dụng PDF đại diện cho ảnh từ các đặc trưng được trích xuất cũng được đề nghị trong những năm gần đây. Tuy nhiên, hầu như chúng chỉ được áp dụng cho bài toán nhận dạng chưa được giám sát (Thao et al., 2023). Với nhận dạng được giám sát, việc đưa ảnh về PDF đại diện để thực hiện còn rất hạn chế.

Nghiên cứu này xây dựng thuật toán phân loại cho PDF dựa trên độ đo được gọi là hệ số chồng lấp và nguyên tắc phân loại tựa Bayes. Thuật toán đề nghị sau đó được áp dụng cho dữ liệu ảnh khi các đặc trưng màu sắc của nó được trích xuất và biểu diễn thành các PDF một chiều. Ý tưởng nghiên cứu này rất thú vị, có thể áp dụng cho nhiều vấn đề thực tế liên quan đến tự động hoá và trí tuệ nhân tạo đang được rất nhiều nhà khoa học quan tâm hiện nay.

2. THUẬT TOÁN PHÂN LOẠI HÀM MẬT ĐỘ XÁC SUẤT ĐỀ NGHỊ

2.1. Độ đo đánh giá sự tương tự của hai hàm mật độ xác suất

Trong không gian xác suất $(\Omega, \mathcal{F}, \mathbb{P})$, với Ω là không gian mẫu, họ $\mathcal{F}$ các tập con đo được của σ-đại số trong Ω và hàm xác suất $\mathbb{P}: \mathcal{F} \to [0; 1]$ gồm hai PDF f và g. Khi đó, ta có các độ đo sau:

– Khoảng cách $\mathcal{L}^1$:

$$\|f, g\|_1 = \int_{\mathbb{R}^n} |f(x) - g(x)|\, dx.$$

– Khoảng cách Divergence:

$$\mathcal{D}_{DVG}(f \parallel g) = 2 \int_{-\infty}^{+\infty} \frac{(f(x) - g(x))^2}{(f(x) + g(x))^2} dx.$$

– Khoảng cách Kullback-Leibler:

$$\mathcal{D}_{KL}(f \parallel g) = \int_{-\infty}^{+\infty} f(x). \ln\left(\frac{f(x)}{g(x)}\right) dx.$$

– Khoảng cách Jensen-Shannon:

$$\mathcal{D}_{JS}(f \parallel g) = \frac{1}{2}\mathcal{D}_{KL}(f \parallel h) + \frac{1}{2}\mathcal{D}_{KL}(g \parallel h),$$

trong đó $h = \frac{1}{2}(f + g)$.

– Khoảng cách Bhattacharyya:

$$\mathcal{D}_B(f,g) = \int_{\mathbb{R}^n} [f(x)g(x)]^{\frac{1}{2}}\,dx.$$

Các khoảng cách trên được sử dụng để đánh giá sự tương tự của hai PDF. Khi khoảng cách càng nhỏ thì sự tương tự của chúng càng lớn và ngược lại. Có nhiều nghiên cứu việc chọn khoảng cách tối ưu cho bài toán nhận dạng thống kê. Tuy nhiên, chưa có khoảng cách nào được xem là tốt nhất cho mọi trường hợp.

Các khoảng cách trên có giá trị trong [0; +∞). Trong bài toán phân loại, chúng ta cần một độ đo đánh giá sự tương tự của 2 PDF nhưng có giá trị [0;1]. Trong nghiên cứu này, một độ đo được gọi là hệ số chồng lấp của 2 PDF được sử dụng. Độ đo này được định nghĩa như sau:

$$D_O(f,g) = 2 - \int_{R^n} \max\{f,g\}dx. \qquad (1)$$

Từ (1) ta có thể nhận thấy $D_O(f,g)$ là thước đo mức độ chồng lấp giữa hai hàm PDF. Khi hai PDF f và g càng có nhiều vùng chồng lấp thì $\int_{R^n} \max\{f,g\}dx$ sẽ càng nhỏ, do đó $D_O(f,g)$ sẽ càng lớn. Ngược lại khi vùng chồng lấp của f và g càng nhỏ thì $\int_{R^n} \max\{f,g\}dx$ sẽ càng lớn, do đó $D_O(f,g)$ sẽ càng nhỏ. Giá trị của $D_O(f,g)$ là 1 khi hai PDF trùng nhau và giá trị của $D_O(f,g)$ là 0 khi hai PDF hoàn toàn tách rời nhau. Từ các nhận xét trên, ta thấy hệ số chồng lấp là độ đo phù hợp với mục tiêu trong xây dựng bài toán phân loại.

2.2. Thuật toán

Cho k tổng thể $w_1, w_2, \dots, w_k$, trong đó nhóm thứ i có n_i PDF, $n_1 + n_2 + \cdots + n_k = N$. Để phân loại PDF f_0, ta thực hiện theo các bước sau:

Bước 1. Thiết lập ma trận xác suất ban đầu $\boldsymbol{U}^{(0)}$ với k hàng và $N+1$ cột:

$$\boldsymbol{U}^{(0)} = [\mu_{ij}]_{k\times(N+1)} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1N} & 1/k \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2N} & 1/k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{kN} & 1/k \end{bmatrix},$$

trong đó N cột đầu tiên là ma trận phân vùng không mờ của các phần tử trong tập dữ liệu khi xếp vào k nhóm $w_1, w_2, \dots, w_k$. Cụ thể $\mu_{ij} = 1$ khi phần tử thứ j thuộc chùm thứ i và $\mu_{ij} = 0$ nếu phần tử thứ j không thuộc chùm thứ i. Cột cuối cùng $(N+1)$ là xác suất ban đầu để f_0 xếp vào các chùm $w_1, w_2, \dots, w_k$. Ở bước này, ta cho xác suất tiên nghiệm của f_0 bằng nhau và bằng $1/k$.

Bước 2. Thiết lập PDF đại diện cho các chùm w_i theo công thức sau:

$$f_{v_i} = \frac{1}{\sum_{j=1}^{N+1}(\mu_{ij})^2} \times \sum_{j=1}^{N+1} (\mu_{ij})^2 f_j, i = 1, 2, \dots, k, \quad (2)$$

trong đó μ_{ij} là xác suất xếp hàm mật độ f_j vào chùm w_i.

Dựa vào Công thức (2), ta dễ dàng nhận được hàm đại diện nhóm f_{v_i} không âm và $\int_{\mathbb{R}^d} f_{v_i}(x)\mathrm{d}x = 1$. Như vậy hàm đại diện của một nhóm cũng là một PDF.

Bước 3. Cập nhật ma trận xác suất mới $\boldsymbol{U}^{(t)}$, trong đó các phần tử được xác định theo công thức sau:

$$\mu_{ij}^{(t)} = \frac{1}{\sum_{m=1}^{N+1}\left[D_O\left(f_{v_i}, f_j\right)/\left(D_O\left(f_{v_i}, f_m\right)\right)^2\right]}, \qquad (3)$$

trong đó $D_O(.)$ là hệ số chồng lấp của các PDF.

Bước 4. Tính chuẩn

$$\left\|\boldsymbol{U}^{(t)} - \boldsymbol{U}^{(t-1)}\right\| = \max_{i,j}\left\{\left|\mu_{ij}^{(t)} - \mu_{ij}^{(t-1)}\right|\right\}.$$

Bước 5. Lặp lại các Bước 2, Bước 3 và Bước 4 t lần cho đến khi điều kiện sau được thỏa:

$$\left\|\boldsymbol{U}^{(t)} - \boldsymbol{U}^{(t-1)}\right\| < \varepsilon.$$

Khi thuật toán kết thúc, ta nhận được một ma trận cỡ $k \times (N+1)$ mà tổng các cột của nó luôn bằng 1. Cột cuối cùng trong ma trận này chính là xác suất tiên nghiệm để xếp PDF f_0 vào các nhóm.

Bước 6. Giả sử ta có cột cuối cùng trong ma trận khi kết thúc Bước 5 là

$$q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix}.$$

Tính $g_{i0} = q_i D_O(f_0, f_{v_i})$ và thực hiện nguyên tắc phân loại như sau:

Nếu $\max\left\{q_i\left(D_O(f_0, f_{v_i})\right)\right\} = q_c\left(D_O(f_0, f_{v_c})\right)$, $c = 1, 2, \dots, k$ thì PDF f_0 được xếp vào nhóm w_c.

Thuật toán phân loại cho các PDF trên có hai giai đoạn chính. Giai đoạn thứ nhất là tìm xác suất tiên nghiệm cho PDF f_0. Trong nhiều nghiên cứu, xác suất tiên nghiệm thường chỉ dựa vào tập huấn luyện

mà không xem xét đến phần tử phân loại. Trong thuật toán này, mối quan hệ mờ giữa PDF được phân loại với các nhóm qua bài toán phân tích chùm được sử dụng làm xác suất tiên nghiệm. Giai đoạn thứ hai của thuật toán phân loại đề nghị là xây dựng nguyên tắc phân loại. Một PDF được xếp vào một nhóm cụ thể nào đó nếu nó có xác suất tiên nghiệm lớn nhất và sự tương tự đến nhóm đó cũng lớn nhất. Do đó, nguyên tắc phân loại này được xem là tựa phương pháp Bayes cho biến rời rạc.

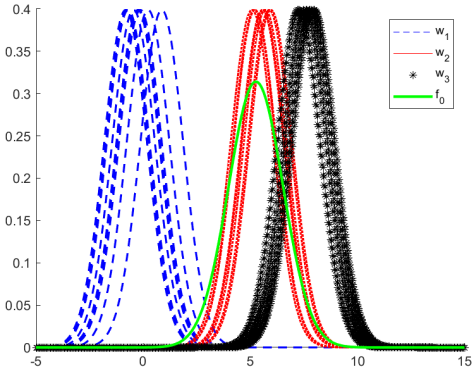
2.3. Ví dụ minh hoạ

Để minh họa thuật toán đề nghị, 30 PDF được xem xét. Các PDF này được chia thành 3 nhóm w_1,w_2 và w_3, với 10 PDF cho mỗi nhóm. Mỗi PDF tuân theo phân phối chuẩn một chiều với phương sai là 1. Tham số trung bình của các nhóm cụ thể được cho như sau:

- $\mu_1 = \{-0{,}82;\ -0{,}74; -0{,}70; -0{,}45;$ $-0{,}25; -0{,}22;\ -0{,}21;\ -0{,}13;\ 0{,}23; 0{,}87\}$,
- $\mu_2 = \{5{,}08;\ 5{,}19;\ 5{,}22;\ 5{,}56;\ 5{,}58;\ 5{,}60;$ $5{,}70;\ 5{,}88;\ 5{,}96\}$,
- $\mu_3 = \{7{,}24;\ 7{,}28;\ 7{,}50;\ 7{,}60;\ 7{,}62;\ 7{,}75;$ $7{,}84;\ 7{,}85;\ \ 7{,}95;\ 7{,}97\}$.

Giả sử f_0 là PDF được phân loại có phân phối chuẩn $N(5.26,1.27)$.

Đồ thị của các PDF cho 3 nhóm và f_0 được cho bởi Hình 1.



Hình 1. Đồ thị PDF của 3 nhóm w_1, w_2, w_3 và f_0

Từ Hình 1, ta có nhận xét rằng f_0 gần nhóm w_2 nhất nên việc phân loại nó vào nhóm này là phù hợp nhất.

Các bước của thuật toán đề nghị được minh hoạ trên tập dữ liệu này như sau:

Bước 1. Thiết lập ma trận phân vùng ban đầu có kích thước 3×31 như sau:

$$\boldsymbol{u}_{3\times 31}^{(0)} = \begin{bmatrix} 1 & 1 & \dots & 0 & 0 & \dots & 0 & 0 & 1/3 \\ 0 & 0 & \dots & 1 & 1 & \dots & 0 & 0 & 1/3 \\ 0 & 0 & \dots & 0 & 0 & \dots & 1 & 1 & 1/3 \end{bmatrix}.$$

Trong ma trận trên, ta có

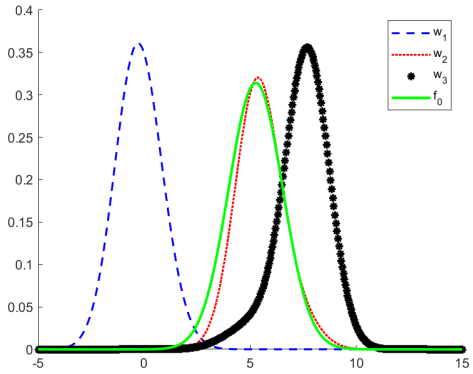
* Hàng 1: Từ cột 1 đến cột 10 là số 1, cột 11 đến cột 30 là số 0 và cột 31 là 1/3.

* Hàng 2: Từ cột 11 đến cột 20 là số 1, cột 31 là 1/3 và các cột còn lại đều là số 0.

* Hàng 3: Từ cột 21 đến cột 30 là số 1, cột 31 là 1/3 và các cột còn lại đều là số 0.

Cột cuối cùng là 1/3, nghĩa là lúc đầu ta cho xác suất để xếp f_0 và ba nhóm giống nhau.

Bước 2. Thiết lập 3 PDF đại diện cho 3 nhóm, ta có Hình 2.



Hình 2. Đồ thị của các PDF đại diện và PDF được phân loại

Bước 3. Cập nhật ma trận phân vùng, ta có

$$\boldsymbol{u}_{3\times 31}^{(0)} = \begin{bmatrix} 0{,}361 & 0{,}361 & \dots & 0{,}315 & 0{,}314 & \dots & 0{,}322 & 0{,}322 & 0{,}308 \\ 0{,}328 & 0{,}329 & \dots & 0{,}311 & 0{,}308 & \dots & 0{,}335 & 0{,}336 & 0{,}306 \\ 0{,}311 & 0{,}310 & \dots & 0{,}374 & 0{,}378 & \dots & 0{,}343 & 0{,}341 & 0{,}386 \end{bmatrix}.$$

Bước 4. Vì $\left\|\boldsymbol{u}^{(1)} - \boldsymbol{u}^{(0)}\right\| = 0.078 > 0.01$ nên thuật toán tiếp tục được lặp lại.

Bước 5. Lặp lại Bước 3 và Bước 4, sau 8 vòng lặp, ta nhận được ma trận sau:

$$\boldsymbol{u}_{3\times 31}^{(8)} = \begin{bmatrix} 0{,}895 & 0{,}922 & \dots & 0{,}042 & 0{,}025 & \dots & 0{,}041 & 0{,}035 & 0{,}023 \\ 0{,}049 & 0{,}033 & \dots & 0{,}907 & 0{,}943 & \dots & 0{,}078 & 0{,}064 & 0{,}943 \\ 0{,}056 & 0{,}045 & \dots & 0{,}051 & 0{,}042 & \dots & 0{,}881 & 0{,}901 & 0{,}034 \end{bmatrix}.$$

Vì $\left\|\boldsymbol{u}^{(8)} - \boldsymbol{u}^{(7)}\right\| = 0.008 < 0.01$ nên Bước 4 dừng lại.

Bước 6. Từ cột cuối cùng trong ma trận Bước 4, chúng ta nhận được xác suất tiên nghiệm để xếp f_0 là

$$q = \begin{bmatrix} 0{,}023 \\ 0{,}943 \\ 0{,}034 \end{bmatrix}.$$

Ta có

$D_O(f_0, f_{v_1}) = 0{,}022$; $D_O(f_0, f_{v_2}) = 0{,}858$;
$D_O(f_0, f_{v3}) = 0{,}355$; $\mu_1 D_O(f_0, f_{v_1}) = 0{,}00051$;
$\mu_2 D_O(f_0, f_{v_2}) = 0{,}809$; $\mu_3 D_O(f_0, f_{v_3}) = 0{,}012$.

Vì $\max\{\mu_i D_O(f_0, f_{vi})\} = \mu_2 D_O(f_0, f_{v2})$ nên f_0 được phân loại vào w_2. Từ Hình 1, ta thấy rằng đây là kết quả phân loại phù hợp.

3. ÁP DỤNG TRONG PHÂN LOẠI ẢNH

3.1. Không gian màu của ảnh

Không gian màu RGB là một hệ thống biểu diễn màu sắc bằng cách kết hợp ba màu cơ bản là đỏ (red), xanh lục (green) và xanh lam (blue) với các tỷ lệ khác nhau. Không gian màu RGB được sử dụng rộng rãi trong các thiết bị điện tử như máy tính, máy ảnh, điện thoại và màn hình hiển thị, vì nó gần với cách mắt người nhận biết màu sắc. Mỗi màu trong không gian màu RGB có thể có giá trị từ 0 đến 255, tương ứng với 8 bit. Do đó, có tổng cộng 256 × 256 × 256 = 16.777.216 màu khác nhau có thể được tạo ra trong không gian màu RGB.

Ảnh xám hay còn gọi là ảnh đơn sắc (monochromatic). Ảnh 8 mức xám mỗi điểm ảnh sẽ có giá trị nằm trong đoạn [0;7], ảnh 256 mức xám mỗi điểm ảnh sẽ có giá trị nằm trong đoạn [0,255]. Giá trị của điểm ảnh bằng 0 đại diện cho điểm ảnh tối (đen), giá trị điểm ảnh lớn nhất đại diện cho điểm ánh sáng (trắng). Độ sáng được tính theo công thức (chuyển đổi từ hệ màu RGB):

$$S = 0.2989R + 0.5870G + 0.1140B.$$

3.2. Phương pháp trích xuất hình ảnh

Điểm ảnh là một phần tử của hình ảnh kỹ thuật số tại một tọa độ với độ xám hoặc màu nhất định. Cụ thể, trong hình ảnh hai chiều, mỗi điểm ảnh bao gồm một cặp tọa độ (x, y), trong đó x, y là các giá trị đại diện cho độ sáng cụ thể (mức xám). Các cặp tọa độ (x, y) tạo nên độ phân giải của ảnh. Cụ thể, màn hình máy tính có độ phân giải 768×1024 có nghĩa chiều rộng màn hình là 768 pixel và chiều dài là 1024 pixel.

Trong nghiên cứu này, việc trích xuất hình ảnh thành PDF dựa trên độ phân giải của ảnh theo các bước sau:

Bước 1. Xây dựng ma trận điểm ảnh với kích thước của mỗi hình ảnh trong bộ dữ liệu.

Bước 2. Chuyển ma trận điểm ảnh thành véctơ cột.

Bước 3. Ước lượng phi tham số mật độ điểm ảnh dựa trên phương pháp hàm hạt nhân.

Giả sử ta có dữ liệu rời rạc d-chiều, khi đó PDF được ước lượng theo phương pháp hàm hạt nhân có dạng:

$$f(x) = \frac{1}{N} \cdot \frac{1}{h_1 h_2 \dots h_d} \sum_{i=1}^{N} \prod_{j=1}^{d} K_j \left(\frac{x_j - x_{ij}}{h_j} \right),$$

trong đó

N là số phần tử của dữ liệu,

d là số chiều,

h_j là tham số trơn của biến thức j,

x_j là biến thứ $j, j = 1,2, \dots, d$,

x_{ij} là dữ liệu thứ i của biến thứ $j, i = 1,2, \dots, N$,

$K_j(\cdot)$ là hàm hạt nhân của biến thứ j. Hàm hạt nhân phải thỏa mãn hai điều kiện $K_j(x) \geq 0$ và $\int K_j(x)\, dx = 1$.

Trong nghiên cứu này chúng tôi chọn hàm mật độ dạng chuẩn:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right).$$

Kết quả ước lượng PDF bằng phương pháp hạt nhân cũng phụ thuộc vào tham số trơn. Cũng có nhiều thảo luận khác nhau về việc chọn tham số trơn, tuy nhiên chưa có phương pháp nào được xem là tối ưu. Trong nghiên cứu này, tham số trơn được chọn theo Terrell (1992).

Bước 4. Áp dụng thuật toán phân loại cho PDF được trình bày Phần 2.2.

4. ỨNG DỤNG TRONG PHÂN LOẠI ẢNH

4.1. Tham số đánh giá các thuật toán phân loại

Để đánh giá sự hiệu quả của mô hình phân loại, ta thường dùng các tham số như Accuracy (*ACC*), Precision, Recall và F1 score. Các tham số này đều dựa vào ma trận "Confusion":

Bảng 1. Mô tả ma trận Confusion trong phân loại

Nhãn nguồn	Nhãn phân loại	
	Lớp 1	Lớp 2
Lớp 1	TP	FP
Lớp 2	FN	TN

Trong ma trận Confusion, ta có

– TP: Tổng số phần tử của Lớp 1 được phân loại đúng

– TN: Tổng số phần tử của Lớp 2 được phân loại đúng

– FP: Tổng số phần tử thuộc Lớp 1 phân loại sai

– FN: Tổng số phần tử thuộc Lớp 2 phân loại sai

Gọi TS là tổng số phần tử được phân loại, khi đó ta có công thức tính các tham số đánh giá như sau:

– ACC = $\frac{TP+TN}{TS}$

– Recall = $\frac{TP}{TP+FN}$

– Precision = $\frac{TP}{TP+FP}$

– F1-Score = $\frac{2\ \text{Precison.Rcall}}{2\ \text{Precison} + \text{Rcall}}$

Các tham số trên có giá trị [0;1]. Khi giá trị này càng lớn thì kết quả phân loại càng tốt và ngược lại.

4.2. Dữ liệu và phương pháp

Phần này áp dụng cụ thuật toán đề nghị để phân loại cho tập ảnh 3 loại ảnh trái cây. Bộ số liệu gồm 1470 ảnh, trong đó có 490 ảnh quả lựu (w_1), 490 ảnh quả táo (w_2) và 490 ảnh quả vải thiều (w_3). Số liệu được trích từ tập dữ liệu Fruit 360 trên Website: https://www.kaggle.com/datasets/ moltean/fruits. Ảnh minh hoạ cho ba nhóm được cho bởi Hình 3.

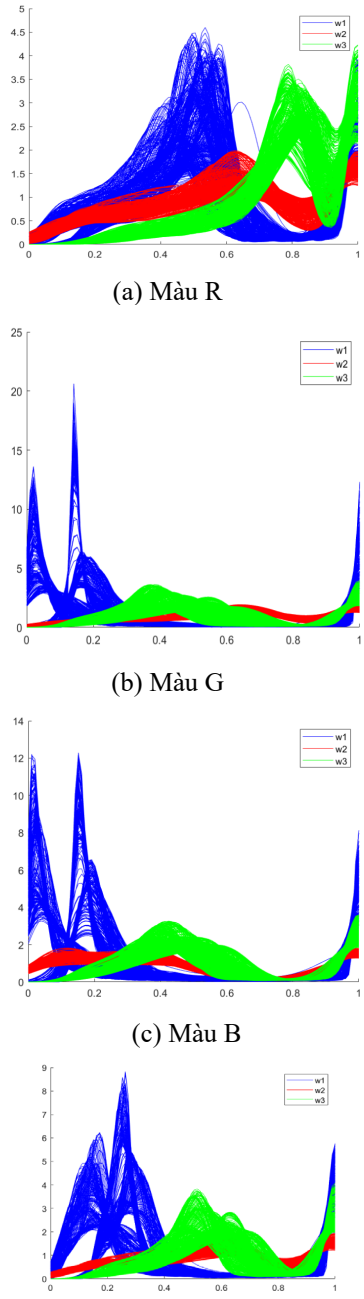
(a) **Lựu** **(b) Táo** **(c) Vải thiều**

Hình 3. Mẫu ảnh của 3 loại trái cây

Với ứng dụng này, nghiên cứu sử dụng 85% của mỗi nhóm ảnh (1251 ảnh của 3 nhóm và 417 ảnh của mỗi nhóm) làm tập huấn luyện để xây dựng thuật toán đề nghị và 15% ảnh còn lại (219 ảnh của 3 nhóm và 73 ảnh của mỗi nhóm) làm tập kiểm tra.

4.3. Kết quả thực hiện

Trích xuất đặc trưng màu sắc của tất cả các ảnh theo 4 màu R, G, B, Gr và ước lượng PDF cho 3 nhóm trái cây, ta nhận được Hình 4.



(a) Màu R

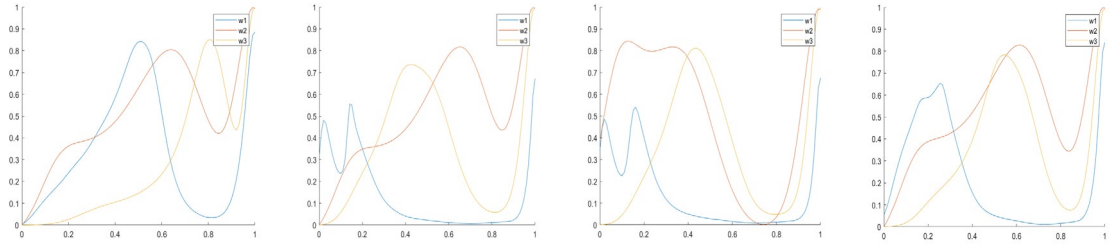
(b) Màu G

(c) Màu B

(d) Màu Gr

Hình 4. Các PDF đại diện cho đặc trưng màu sắc được trích xuất từ 1470 ảnh.

Tìm PDF đại diện cho 3 nhóm trái cây từ 4 đặc trưng màu sắc được trích xuất, ta có Hình 5.



(a) Màu R (b) Màu G (c) Màu B (c) Màu Gr

Hình 5. Các PDF đại diện cho 3 nhóm trái cây được trích xuất đặc trưng từ 4 màu

Sử dụng các PDF được trích xuất cho từng ảnh, PDF đại diện cho mỗi nhóm và áp dụng thuật toán trong các trường hợp khác nhau, ta nhận được Bảng 2.

Bảng 2. So sánh kết quả phân loại của thuật toán đề nghị với các khoảng cách khác nhau

Màu	Số vòng lặp	ACC	Precison	F1-score
R	23	0,7714	0,7660	0,7660
G	23	0,8048	0,7714	0,7714
B	23	0,7660	0,8048	0,8048
Gr	24	0,8150	0,8150	0,8150

Từ Bảng 2, ta có những nhận xét sau:

– Giá trị $ACC = 0{,}7714$ của phân loại khi các đặc trưng được trích xuất từ màu R là thấp nhất, trong khi đặc trưng được trích xuất từ màu Gr cho giá trị $ACC = 0{,}8150$ cao nhất.

– Giá trị Precision và F1 của mỗi trường hợp là bằng nhau. Chúng ta một lần nữa cũng nhận Precision và F1 thấp nhất khi sử dụng màu R và giá trị cao nhất khi sử dụng đặc trưng màu Gr trong phân loại.

– Khi sử dụng các đặc trưng được trích xuất từ màu G, B và R, số vòng lặp thực hiện gần như nhau (23 hoặc 24 vòng lặp). Số vòng lặp thực hiện khi sử dụng đặc trưng màu Gr nhiều nhất và có sự chênh lệch với các trường hợp còn lại. Do đó thời gian thực hiện khi sử dụng màu Gr là nhiều nhất.

Như vậy trong tập dữ liệu này, việc sử dụng đặc trưng màu Gr để thực hiện thuật toán đề nghị sẽ nhận được kết quả phân loại tốt nhất. Mặc dù trường hợp này thời gian thực hiện lâu hơn các trường hợp khác, nhưng với sự phát triển mạnh mẽ trong tốc độ xử lý tính toán của các loại máy tính hiện nay, vấn đề này cũng không ảnh hưởng nhiều đến hiệu quả thực tế.

5. KẾT LUẬN

Nghiên cứu này đã đề xuất được một thuật toán phân loại cải tiến cho PDF với sự trình bày chi tiết các bước thực hiện và minh hoạ cụ thể. Một vấn đề quan trọng của nghiên cứu này là việc vận dụng thuật toán đề nghị cho dữ liệu ảnh khi đặc trưng màu sắc của nó được trích xuất và biểu diễn thành các PDF. Ý tưởng thú vị này có sự khác biệt so với các phương pháp thông dụng hiện nay khi đặc trưng của ảnh chủ yếu được biểu diễn thành ma trận hoặc vectơ đại diện. Thuật toán đề nghị có thể vận dụng trong thực tế của nhiều lĩnh vực khác nhau để cải tiến hiệu quả của sự phân loại.

Trong thời gian sắp tới, các ứng dụng thực tế cho ảnh sẽ được tiếp tục thực hiện để đánh giá sự hiệu quả của thuật toán đề nghị so với các thuật toán khác. Hơn nữa, thuật toán phân loại ảnh tiếp tục được mở rộng khi cùng lúc nhiều đặc tính của chúng được trích xuất và được biểu diễn bởi nhiều PDF một chiều hoặc một PDF nhiều chiều để nhận diện.

LỜI CẢM ƠN

Đề tài này được tài trợ bởi Trường Đại học Cần Thơ, Mã số: TSV2024-54.

TÀI LIỆU THAM KHẢO

Cinar, I. A. (2022). Identification of rice varieties using machine learning algorithms. *Journal of Agricultural Sciences, 28*(2), 307-325. https://doi.org/10.15832/ankutbd.862482

Che-Ngoc, H., Nguyen-Trang, T., Huynh-Van, H. (2023). Improving Bayesian classifier using vine copula and fuzzy clustering Technique. *Annal of Data Science, 11*, 709–732. https://doi.org/10.1007/s40745-023-00490-4

Dan, N. T. & Tai, V. V. (2024). Classifying for interval and applying for image based on the extracted texture feature. *Granular Computing, 9*(29), 1-18. https://doi.org/10.1007/s41066-024-00450-0

Dinh, P. T. & Tai, V. V. (2021). Automatic fuzzy genetic algorithm in clustering for images based on the extracted intervals. *Multimedia Tools and Applications, 80*(28), 35193-35215. https://doi.org/10.1007/s11042-020-09975-3

Dinh, P. T. & Tai, V. V. (2023). Improving the genetic algorithm in fuzzy cluster analysis for numerical data and its applications. *Iranian Journal of Fuzzy Systems*, *20*(5), 171-187. https://doi: 10.22111/ijfs.2023.7834

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics, 8*(4), 376–386. https://doi.org/10.1007/s11042-020-09975-3

Ha, C. N., Thao, N. T., Bao, T. N. Trung, N. T. & Tai, V. V. (2020). A new approach for face detection using the maximum function of probability density functions. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03823-1

Huang, S., N., Cai, P., Pacheco, P., Narrandes, S., Wang, Y. & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer geometrics. *Cancer Genomics-Proteomics, 15*(1), 41–51. https:// doi.org/10.21873/cgp.20063

Imandoust, S. B. & Bolandraftar, M. (2013). Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications, 3*(5), 605-610.

Katarzyna, S. (2017). Evaluation of classifiers: current methods and future research directions. *Computer Science and Information Systems*, *12*, 37–40. https://doi.org/10.15439/2017F530 ISSN 2300-5963

Koklu, M. C. (2021). Classification of rice varieties with deep learning methods. *Computers and Electronics in Agriculture, 187*, 106285. https://doi.org/10.1016/j.compag.2021.106285

Ngoc, L. T. K., Tuan, L. H. & Tai, V. V. (2021). Automatic clustering algorithm for interval data based on overlap distance. *Communications in Statistics - Simulation and Computation*. https://doi.org//10.1080/03610918.2021.1900248

Ngoc, L.T.K, Thao, N. T. & Tai, V. V. (2022). A new image classification method using interval texture feature and improved Bayesian classifier. *Multimedia Tools and Applications.* https://doi.org/10.1007/s11042-022-13531-6

Pham, B. T., Pradhan, B., Bui, D. T., Prakash, I. & Dholakia, M. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of uttarakhand area (India). *Environmental Modelling & Software, 84*, 240–250. https://doi.org/10.1016/j.envsoft.2016.07.005

Pham-Gia, T., Turkkan, N. & Vovan, T. (2000). Statistical discrimination analysis using the maximum function. *Communications in Statistics - Simulation and Computation 37*(2), 320–336. https://doi.org/10.1080/03610910701790475

Philipp, L., Lukas, R., Robert, A. V., Billy J. F., Marius, K. & Klaus-Robert M. (2021). Explainable deep one-class classification. *The conference ICLR 2021*. https://arxiv.org/abs/2007.01760

Tai V. V., Ha, C. N., Nghiep, L. D. & Thao, N. T. (2022). A new strategy for short-term stock in vestment using Bayesian approach. *Computational Economics.* https://doi.org/10.1007/s10614-021- 10115-8

Thao, N. T. & Tai, V. V. (2017). A new approach for determining the prior probabilities in the classification problem by Bayesian method. *Advances in Data Analysis and Classification, 11*(3), 629–643. https://doi.org/10.1080/1351847X.2017.1419273

Thao, N. T., Tai, V. V. & Ha, C. N. (2023). An efficient automatic clustering algorithm for probability density functions and its applications in surface material classification. *Satistica Neerlandica*, *78*(1), 244-260. https://doi.org/10.1111/stan.12315

Zhu, X., Yang, J. & Waibel, A. (2000). Segmenting hands of arbitrary color in automatic face and gesture recognition. *The International Conference on Automatic Face and Gesture Recognition. IEEE.* https://doi.org/ 10.1109/AFGR.2000.84067