



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0026880

(51)⁷ G01S 13/06; G01S 7/02

(13) B

(21) 1-2018-05350

(22) 29/11/2018

(45) 25/01/2021 394

(43) 25/02/2019 371A

(73) Tập đoàn Công nghiệp - Viễn thông Quân Đội (VN)

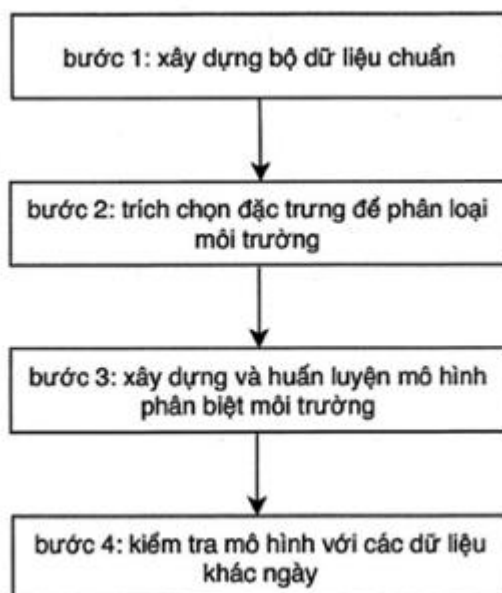
Số 1 đường Trần Hữu Dực, phường Mỹ Đình 2, quận Nam Từ Liêm, thành phố Hà Nội

(72) TRẦN THỊ THANH (VN); LƯU HOÀI NAM (VN); TRẦN QUỐC TUÂN (VN);
TRẦN VŨ HỢP (VN); NGUYỄN VĂN LỢI (VN).

(74) Công ty TNHH Tư vấn Quốc Dân (NACI CO., LTD)

(54) PHƯƠNG PHÁP TỰ ĐỘNG NHẬN DIỆN NHIỀU THỜI TIẾT NHẪM NÂNG CAO KHẢ NĂNG PHÁT HIỆN CHO ĐÀI RAĐA BIÊN

(57) Sáng chế đề cập đến phương pháp giải quyết bài toán tự động nhận diện nhiều thời tiết bằng sự kết hợp giữa thuật toán lựa chọn đặc trưng và kỹ thuật học máy để giải quyết bài toán phân loại môi trường đồng nhất và không đồng nhất. Sáng chế đề xuất thực hiện qua các bước, bước 1: xây dựng bộ dữ liệu chuẩn; bước 2: trích chọn đặc trưng để phân loại môi trường; bước 3: xây dựng và huấn luyện mô hình phân biệt môi trường; bước 4: kiểm tra mô hình với các dữ liệu khác ngày.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp tự động nhận diện nhiều thời tiết nhằm nâng cao khả năng phát hiện cho đài radar biển. Cụ thể, phương pháp tự động nhận diện nhiều thời tiết nhằm nâng cao khả năng phát hiện cho đài radar biển sử dụng học máy để giải quyết bài toán nhận diện nhiều thời tiết trong lĩnh vực xử lý nhiều tín hiệu, tăng khả năng phát hiện cho đài radar biển, đảm bảo an ninh biển.

Tình trạng kỹ thuật của sáng chế

Hiện nay, vấn đề tự động nhận diện nhiều thời tiết đóng vai trò rất quan trọng trong lĩnh vực xử lý nhiều tín hiệu radar. Trong quá trình giải quyết vấn đề, đã có nhiều khó khăn và thách thức được đặt ra. Trong đó, có thể kể đến sự thay đổi thường xuyên của môi trường và thời tiết, điều này ảnh hưởng rất lớn đến khả năng phát hiện mục tiêu, do đặc điểm đài có độ nhạy lớn đối với các đặc trưng biến đổi của thời tiết. Những phương pháp loại nhiễu truyền thống có hạn chế lớn là ngưỡng bù thay đổi tùy theo điều kiện cụ thể về thời tiết, đòi hỏi trắc thủ phải quan sát và linh động trong việc điều chỉnh thủ công giá trị ngưỡng bù đó.

Thông thường, việc tự động nhận diện nhiều thời tiết sử dụng phương pháp xác định ngưỡng thích nghi môi trường, nguyên lý cơ bản của thuật toán này là tính toán và so sánh độ chênh lệch giá trị công suất Δd giữa các ô với một giá trị C , nếu số lượng Δd vượt ngưỡng C lớn hơn 1, tức môi trường không đồng nhất, ta sẽ sử dụng thuật toán ổn định xác suất báo động làm thống kê theo thứ tự, ngược lại sẽ sử dụng thuật toán ổn định xác suất báo động làm trung bình giá trị các ô. Sơ đồ khối của phương pháp xác định ngưỡng thích nghi môi trường truyền thống được thể hiện trong Hình 1.

Nhược điểm của phương pháp phân loại môi trường dựa vào thuật toán thích nghi môi trường trên là việc xác định được giá trị C hợp lý. Hơn nữa, giá trị C này cũng sẽ thay đổi theo điều kiện thay đổi của thời tiết.

Ngày nay, với sự phát triển không ngừng của khoa học dữ liệu, trí tuệ nhân tạo đã phân nào giúp cho phương pháp phân loại môi trường trở nên dễ dàng hơn. Xuất phát từ hạn chế của các phương pháp đã biết, nhóm tác giả đã đề xuất một giải pháp sử dụng học máy để giải quyết vấn đề tự động nhận diện nhiều thời tiết.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp tự động nhận diện nhiều thời tiết nhằm nâng cao khả năng phát hiện cho đài radar biển. Để thực hiện mục đích này, phương pháp được đề xuất bao gồm các bước như sau: bước 1: xây dựng bộ dữ liệu chuẩn; bước 2: trích chọn đặc trưng để phân loại môi trường; bước 3: xây dựng và huấn luyện mô hình phân biệt môi trường; bước 4: kiểm tra mô hình với các dữ liệu khác ngày.

Phương pháp tự động nhận diện nhiều thời tiết được sử dụng để phân loại ra hai loại môi trường cụ thể: đồng nhất và không đồng nhất. Kết quả sau đó sẽ được đưa vào bộ phát hiện mục tiêu sử dụng thuật toán ổn định xác suất báo động lầm, nếu môi trường đồng nhất sẽ sử dụng thuật toán ổn định xác suất báo động lầm dựa theo trung bình giá trị các ô tham chiếu, nếu không đồng nhất sẽ sử dụng thuật toán ổn định xác suất báo động lầm dựa vào thống kê theo thứ tự giá trị các ô tham chiếu.

Mô tả vắn tắt các hình vẽ

Hình 1 là hình vẽ thể hiện sơ đồ thuật toán xác định ngưỡng thích nghi môi trường truyền thống;

Hình 2 là hình vẽ thể hiện phương pháp giải quyết bài toán tự động nhận diện nhiều thời tiết;

Hình 3 là hình vẽ thể hiện thuật toán trích xuất lựa chọn đặc trưng;

Hình 4 là hình vẽ thể hiện mô hình kỹ thuật đóng gói được sử dụng;

Hình 5 là hình vẽ thể hiện độ chính xác phân loại theo số lượng tập dữ liệu con được tạo ra.

Mô tả chi tiết sáng chế

Sáng chế đề xuất phương pháp giải quyết bài toán phân loại nhiều thời tiết có các bước thực hiện như tham chiếu tại Hình 2. Chi tiết từng bước của phương pháp được trình bày cụ thể dưới đây:

Bước 1: xây dựng bộ dữ liệu chuẩn.

Do đặc điểm đài radar có độ nhạy lớn đối với các đặc trưng thay đổi của thời tiết, do vậy dữ liệu sẽ có sự khác biệt đáng kể giữa điều kiện thời tiết tốt (đồng nhất) và điều kiện thời tiết xấu (không đồng nhất). Nhóm tác giả đã trích xuất dữ liệu từ hai môi trường: môi trường đồng nhất thu được từ bộ dữ liệu ghi lưu trong điều kiện thời tiết tốt, trời quang, ít mây; môi trường không đồng nhất thu được từ bộ dữ liệu ghi lưu trong điều kiện thời tiết xấu, mưa, gió, mây mù. Để thu được bộ dữ liệu có chất lượng - thể hiện rõ nhất đặc tính của hai loại môi trường - được sử dụng làm bộ dữ liệu chuẩn - bộ dữ liệu thời tiết đặc trưng tại một vùng nhất định - để xây dựng mô hình dự đoán, nhóm tác giả đã quan sát dữ liệu hiển thị trên màn hình video để phân loại được rõ dữ liệu từ hai loại môi trường này. Để đảm bảo tốc độ xử lý thời gian thực và không bị lẫn mục tiêu vào môi trường, tạo một cửa sổ trượt với tổng số các ô tham chiếu n trong mỗi cửa sổ là khoảng từ 1200 đến 1300. Số lượng cửa sổ - số lượng mẫu dùng cho huấn luyện mô hình được sử dụng khoảng một triệu. Các ô tham chiếu trong mỗi cửa sổ sẽ được trích xuất các đặc trưng của chúng.

Bước 2: trích chọn đặc trưng để phân loại môi trường.

Tham chiếu Hình 2, ý tưởng cơ bản của phương pháp là đưa dữ liệu về hình dạng một phân bố, dạng phân bố này là khác nhau đối với mỗi loại môi trường. Đối với các loại dữ liệu khác nhau, phương pháp đưa dữ liệu về hình dạng một phân bố là khác nhau. Đối với dữ liệu radar cảnh giới bờ tầm trung, tham chiếu vào Hình 3, bước quan trọng ban đầu trong phương pháp là phải xác định được số điểm k (số ngăn) để xây dựng được dạng phân bố, để giải quyết vấn đề này chúng tôi sử dụng định luật Rice, theo đó, số điểm k được sử dụng để xây dựng nên dạng phân bố được tính như sau:

$$k = 2 * n^{1/3}$$

Giá trị k sau đó sẽ được làm tròn đến lũy thừa hai tiếp theo. Với số lượng điểm k đã xác định, tiến hành tính độ rộng Δw giữa mỗi điểm theo công thức sau:

$$\Delta w = \frac{x_{max} - x_{min}}{k}$$

Trong đó, x_{max} và x_{min} tương ứng là giá trị lớn nhất và giá trị nhỏ nhất trong n ô tham chiếu, độ rộng Δw sẽ được sử dụng cho việc đếm số ô trong n ô tham chiếu có giá trị nằm trong khoảng $(x_{min}, x_{min} * \Delta w)$ và các khoảng $(i * x_{min} * \Delta w, (i+1) * x_{min} * \Delta w)$, trong đó i chạy từ 1 đến k .

Tiến hành đếm số lượng các ô trong n ô tham chiếu vào một chuỗi k các khoảng giá trị liên tiếp, từ giá trị x_{min} đến $k * x_{min} * \Delta w$. Sau khi đếm số lượng các ô nằm trong mỗi khoảng, hay chính là đếm tần suất xuất hiện các giá trị nằm trong một khoảng giá trị, vẽ các tần suất xuất hiện trong mỗi khoảng lên một đồ thị (trục ngang là khoảng giá trị, trục đứng là giá trị tần suất) ta sẽ thu được hình dạng một phân bố. Để nâng cao khả năng hoạt động cho các kỹ thuật học máy, k đặc trưng sẽ được chuẩn hóa giá trị theo công thức sau:

$$y_j = \frac{(y_{max} - y_{min}) * (x_j - x_{min})}{x_{max} - x_{min}} + y_{min}, j \in [1, n]$$

Trong đó, y_{max} và y_{min} tương ứng là giá trị lớn nhất và giá trị nhỏ nhất nhất trong khoảng giá trị ta muốn chuẩn hóa, x_{min} và x_{max} tương ứng là giá trị nhỏ nhất và giá trị lớn nhất trong n ô tham chiếu, x_j là giá trị muốn chuẩn hóa, y_j là giá trị sau chuẩn hóa x_j . Bằng việc chuẩn hóa này, tất cả dữ liệu sẽ được đưa về một miền giá trị $[y_{min}, y_{max}]$, đảm bảo sự nhất quán cho dữ liệu.

Bước 3: xây dựng và huấn luyện mô hình học máy phân biệt môi trường.

Tham chiếu Hình 2, để phân loại được môi trường, tiến hành huấn luyện mô hình với hai nhãn “Đồng nhất” và “Không đồng nhất”, sau đó mô hình này sẽ được sử dụng để tự động nhận dạng loại thời tiết trong tương lai.

Để huấn luyện được một mô hình tốt, trước tiên cần phải chuẩn bị được một bộ dữ liệu chuẩn. Các bộ dữ liệu chuẩn này đã được xây dựng tại bước 1, trong đó một nửa dữ liệu được gán nhãn “Đồng nhất”, một nửa được gán nhãn “Không đồng nhất”.

Có rất nhiều kỹ thuật học máy khác nhau được sử dụng để huấn luyện, trong thực tế, các kỹ thuật được sử dụng nhiều là: hồi quy lô-gic, máy véc-tơ hỗ trợ, phân cụm (dùng K hàng xóm gần nhất), cây quyết định, mạng nơ-ron nhân tạo, và các kỹ thuật kết hợp như kỹ thuật đóng gói,... Bản chất của việc học là xác định mô hình (hàm toán học) cho tập “Huấn luyện” được đưa vào học. Mô hình này được xây dựng dựa trên việc tính toán các trọng số. Hồi quy lô-gic là kỹ thuật phân loại nhị phân đơn giản, hàm (mô hình) cần tìm có dạng:

$$y = \begin{cases} 1, & \beta_0 + \beta_1 x + \alpha > 0 \\ 0, & \beta_0 + \beta_1 x + \alpha \leq 0 \end{cases}$$

Trong đó, x là dữ liệu, y là kết quả của mô hình, các tham số β_0 , β_1 , α được tìm trong quá trình học. Ý tưởng cơ bản của máy véc-tơ hỗ trợ là tìm một siêu phẳng có lề (khoảng cách nhỏ nhất giữa bất kỳ đối tượng nào trong tập “Huấn

luyện” và siêu phẳng) cực đại. Về mặt toán học, để tìm siêu phẳng có lề cực đại chúng ta giải quyết bài toán tối ưu dưới đây:

$$\left\{ \begin{array}{l} \text{Cực đại lề } M \\ (\beta_0, \beta_1, \dots, \beta_d) \\ \sum_{j=0}^d \beta_j^2 = 1 \text{ (bài toán tối ưu có ràng buộc)} \\ (\mathbf{Y})^{(i)}(\beta_0 + \beta_1 X_1^{(i)} + \dots + \beta_d X_d^{(i)}) \geq M, \quad \forall i \\ (Y)^{(i)} \in \{-1, 1\} \text{ là các nhãn lớp} \\ d \text{ là số chiều (số đặc trưng) của mỗi mẫu dữ liệu} \end{array} \right.$$

Đối với các dữ liệu mà các lớp không phân tách tuyến tính, sử dụng các hàm Kernel như hàm Kernel đa thức, hàm Kernel Gaussian, hàm Kernel Sigmoid để chuyển dữ liệu sang một không gian mới sao cho dữ liệu ở hai lớp trở nên gần phân biệt tuyến tính. Thuật toán cây quyết định cho ra kết quả là một tập luật của những dữ liệu huấn luyện có thuộc tính (đặc trưng). Cây quyết định là một công cụ phổ biến trong khai phá và phân lớp dữ liệu. Đặc điểm của cây quyết định: là một cây có cấu trúc, trong đó: “Gốc” là nút trên cùng của cây; “Nút trong” là nút trung gian trên một thuộc tính đơn; “Nhánh” biểu diễn các kết quả của kiểm tra trên nút; “Nút lá” biểu diễn lớp hay sự phân phối lớp. Ví dụ ta cần phân loại giữa nam và nữ dựa vào hai thuộc tính chiều cao và cân nặng, ta thực hiện cây quyết định, được xây dựng dựa trên tập luật sau:

Nếu chiều cao $\geq 80\text{cm}$ thì là nam;

Nếu chiều cao $< 80\text{cm}$ và cân nặng $> 60\text{kg}$ thì là nam;

Nếu chiều cao $< 80\text{cm}$ và cân nặng $\leq 60\text{kg}$ thì là nữ;

Kỹ thuật đóng gói là một trong những thuật toán học máy phổ biến nhất và mạnh nhất. Đây là một phương pháp kết hợp sự dự đoán từ nhiều mô hình khác nhau để đưa ra kết quả dự đoán chính xác hơn so với kết quả dự đoán của một mô hình. Thuật toán này được sử dụng để giảm phương sai đối với các thuật toán có

phương sai cao. Thuật toán có phương sai cao là cây quyết định. Các cây quyết định khá nhạy cảm đối với từng tập dữ liệu huấn luyện, đối với một bộ dữ liệu cụ thể, mỗi một cách sắp xếp thứ tự các mẫu sẽ tạo ra một mô hình với độ chính xác khác nhau. Giả sử có một tập dữ liệu 800.000 mẫu và sử dụng thuật toán cây quyết định, kỹ thuật đóng gói của thuật toán cây quyết định hoạt động như sau: bước 1: tạo nhiều (ví dụ 200) tập dữ liệu huấn luyện con từ tập dữ liệu “Huấn luyện” ban đầu, bằng cách thay đổi thứ tự, hoặc lặp lại các mẫu trong tập dữ liệu đó; bước 2: tiến hành luyện mô hình cây quyết định đối với mỗi tập dữ liệu huấn luyện con; bước 3: đối với một tập dữ liệu kiểm tra mới (khoảng 200000 mẫu, không lấy từ tập huấn luyện), tính độ chính xác trung bình từ mỗi mô hình. Ví dụ, giả sử ta có 5 mô hình cây quyết định được đóng gói, đối với một mẫu dữ liệu kiểm tra, ta thu được 5 kết quả từ 5 mô hình như sau: lớp A, lớp A, lớp B, lớp A, lớp A, ta sẽ lấy kết quả xuất hiện nhiều nhất và đưa ra kết quả dự đoán cuối cùng là lớp A. Khi sử dụng kỹ thuật đóng gói các cây quyết định, sẽ tránh được khả năng học vẹt của các cây quyết định (ví dụ như cây quyết định quá sâu, nhiều nút lá...).

Sau khi xây dựng được bộ dữ liệu huấn luyện, nhóm tác giả đã nghiên cứu, và thử nghiệm với các kỹ thuật phân loại khác nhau nêu trên để đưa ra được một kỹ thuật cho kết quả cao nhất đối với dữ liệu thu được từ đài. Quá trình thử nghiệm được thực hiện theo nguyên lý Pareto, khoảng 80% kết quả là do 20% nguyên nhân gây ra, do đó bộ dữ liệu được chia thành thành 2 tập “Huấn luyện” và “Kiểm thử” với tỉ lệ tương ứng là 80:20. Trong đó, tập “Huấn luyện” sẽ được sử dụng để huấn luyện mô hình và tập “Kiểm thử” sẽ được sử dụng để xác định độ chính xác của mô hình đó. Sau đó, với cùng tập “Huấn luyện” và “Kiểm thử” trên, tiến hành thử nghiệm các mô hình với các kỹ thuật như sau:

Mô hình 1: được tạo ra bằng việc sử dụng kỹ thuật véc-tơ hỗ trợ, sử dụng hàm Kernel Gaussian.

Mô hình 2: được tạo ra bằng việc sử dụng kỹ thuật mạng nơ-ron nhân tạo, số lớp ẩn là 1, số nơ-ron ẩn là 10, hàm truyền đạt được sử dụng là Log-Sigmoid, sử dụng giải thuật lan truyền ngược để luyện mạng.

Mô hình 3: được tạo ra bằng việc sử dụng kỹ thuật phân cụm, sử dụng thuật toán K hàng xóm gần nhất, giá trị $K = 30$.

Mô hình 4: được tạo ra bằng việc sử dụng kỹ thuật đóng gói, với 100 tập dữ liệu huấn luyện con, thuật toán học (luyện mô hình) được sử dụng cho mỗi tập con là cây quyết định.

Các giá trị tham số sử dụng trong mỗi kỹ thuật để tạo mô hình đã được nhóm tác giả thử nghiệm trên tập dữ liệu “Huấn luyện”, cho kết quả cao hơn khi sử dụng các giá trị khác. Để đánh giá hiệu quả của mỗi mô hình, sử dụng tập dữ liệu “Kiểm thử”. Mỗi mẫu trong tập “Kiểm thử” sẽ được đưa qua mỗi mô hình, nếu kết quả trả ra là cùng nhãn với nhãn đích (nhãn đã được gán sẵn trong tập dữ liệu “Kiểm thử”) thì kết luận mẫu đó được dự đoán đúng. Sử dụng độ chính xác phân loại A để đánh giá độ chính xác của mỗi mô hình, giá trị A được thực hiện bằng cách tính tỉ lệ giữa số mẫu được dự đoán đúng trên tổng số mẫu trong tập dữ liệu “Kiểm thử”.

$$A = \frac{\text{Số mẫu được dự đoán đúng}}{\text{Tổng số mẫu đưa ra kiểm tra}}$$

Chi tiết về kết quả thực nghiệm được thể hiện ở bảng dưới đây.

Tên mô hình	Độ chính xác A
Mô hình 1	94.6 %
Mô hình 2	95.7 %
Mô hình 3	95.1 %
Mô hình 4	98.7 %

Sau quá trình thử nghiệm, mô hình 4 là mô hình cho kết quả độ chính xác phân loại tốt nhất. Mô hình được minh họa chi tiết trong Hình 4. Quá trình huấn luyện để tạo ra được mô hình 4 sử dụng kỹ thuật đóng gói được thực hiện như sau: tập dữ liệu “Huấn luyện” ban đầu sẽ được chia ra thành 2 tập với tỉ lệ 80: 20, tập lớn dùng để tạo mô hình (“Tập T”), tập nhỏ là tập xác nhận (“Tập X”), tập xác nhận này sẽ được dùng để xác định số tập dữ liệu con cần tạo ra. Bước thứ nhất, tạo ra 20 tập dữ liệu con từ Tập T bằng cách thay đổi vị trí hoặc lặp lại các mẫu dữ liệu. Mỗi tập dữ liệu con sẽ được đưa vào học để tạo ra mô hình bằng thuật toán cây quyết định. Như vậy, với 20 tập dữ liệu con, tương ứng sẽ thu được 20 mô hình. Tập X sẽ được đưa qua từng mô hình, kết quả dự đoán của mỗi mẫu trong Tập X được xác định là kết quả xuất hiện nhiều nhất trong 20 mô hình. Độ chính xác phân loại A của mô hình kết hợp được tính theo công thức nêu trên. Tiếp tục tăng số lượng tập dữ liệu con được tạo ra từ tập T và thực hiện các bước như bước thứ nhất cho đến khi độ chính xác phân loại A bão hòa (không tăng nữa). Hình 5 minh họa quá trình này. Sau khi đã xác định được số tập dữ liệu con cần tạo ra để đạt được độ chính xác phân loại cao, trong trường hợp này là 100, tiến hành đưa toàn bộ dữ liệu trong tập “Huấn luyện” ban đầu vào học bằng cách tạo ra 100 tập dữ liệu con và dùng thuật toán cây quyết định để tạo ra 100 mô hình tương ứng. Tập “Kiểm thử” sẽ được đưa qua từng mô hình, kết quả dự đoán của mỗi mẫu trong tập “Kiểm thử” được xác định là kết quả xuất hiện nhiều nhất trong 100 mô hình. Độ chính xác phân loại của mô hình kết hợp (mô hình 4) được tính theo giá trị A nêu trên.

Bước 4: kiểm tra mô hình với các dữ liệu khác ngày.

Tham chiếu Hình 2, để đánh giá được chính xác hơn về khả năng thực hiện của mô hình 4, nhóm tác giả tiếp tục kiểm tra mô hình với các dữ liệu thu được từ các ngày và các thời tiết khác nhau. Việc thu dữ liệu và xử lý được thực hiện như bước 1 và bước 2. Kết quả độ chính xác phân loại A (theo công thức tại bước 3) thể hiện ở bảng dưới đây.

Tập dữ liệu	A
Ngày 03/03/2018, thời tiết xấu, mây mù theo nhóm	96.61%
Ngày 19/04/2018, thời tiết đẹp, trời đẹp, quang, ít mây	99,08%
Ngày 16/12/2017, thời tiết xấu, mây mù, gió mưa	96,14%

Các kết quả trên cho thấy phương pháp được đề xuất trong sáng chế cho kết quả phân loại tốt, hoàn toàn có thể được sử dụng để nâng cao khả năng phát hiện cho đài radar biển.

Sáng chế được mô tả chi tiết bằng cách sử dụng các phương án được mô tả ở trên. Tuy nhiên, rõ ràng là đối với người hiểu biết trung bình trong lĩnh vực sáng chế không bị giới hạn ở phương án được mô tả trong phần mô tả sáng chế. Sáng chế có thể được thực hiện ở chế độ cải biến hoặc thay đổi mà không nằm ngoài phạm vi sáng chế được xác định bởi các điểm yêu cầu bảo hộ. Vì vậy những gì được mô tả trong phần mô tả sáng chế chỉ nhằm mục đích minh họa, và sẽ không áp đặt bất kỳ giới hạn nào đối với sáng chế.

Yêu cầu bảo hộ

1. Phương pháp tự động nhận diện nhiều thời tiết nhằm nâng cao khả năng phát hiện cho đài radar biên bao gồm bốn bước:

bước 1: xây dựng bộ dữ liệu chuẩn; tại bước này, lấy dữ liệu thực tế quan sát được từ đài cảnh giới bờ tầm trung, dữ liệu được quan sát và phân thành hai loại dữ liệu môi trường đồng nhất và môi trường không đồng nhất; môi trường đồng nhất thu được từ bộ dữ liệu ghi lưu trong điều kiện thời tiết tốt, trời quang, ít mây; môi trường không đồng nhất thu được từ bộ dữ liệu ghi lưu trong điều kiện thời tiết xấu, mưa, gió, mây mù;

bước 2: trích chọn đặc trưng để phân loại môi trường; bước này sử dụng phương pháp phân phối tần suất; số lượng điểm đặc trưng được xác định như sau:

$$k = 2 * n^{1/3}$$

trong đó: k là số điểm được sử dụng để xây dựng nên dạng phân bố, n là các ô tham chiếu trong mỗi cửa sổ;

sau đó tiến hành tính độ rộng Δw giữa mỗi điểm theo công thức sau:

$$\Delta w = \frac{x_{max} - x_{min}}{k}$$

trong đó: x_{max} và x_{min} tương ứng là giá trị lớn nhất và giá trị nhỏ nhất trong n ô tham chiếu; Δw là độ rộng giữa mỗi điểm k ;

tiến hành chia các giá trị của n ô tham chiếu vào một chuỗi k các khoảng giá trị liên tiếp, từ giá trị x_{min} đến $k * x_{min} * \Delta w$, đếm số lượng các ô nằm trong mỗi khoảng;

tạo cửa sổ trượt với tổng số các ô tham chiếu n trong mỗi cửa sổ trong khoảng một nghìn hai trăm đến một nghìn ba trăm sẽ được đưa qua thuật toán phân

phối tần suất để trích xuất các đặc trưng; các đặc trưng được chuẩn hóa theo công thức:

$$y_j = \frac{(y_{max} - y_{min}) * (x_j - x_{min})}{x_{max} - x_{min}} + y_{min};$$

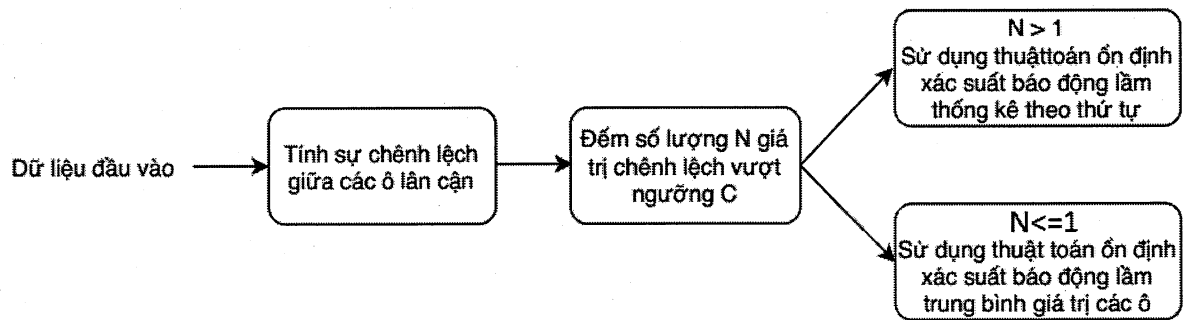
trong đó y_{max} và y_{min} tương ứng là giá trị lớn nhất và giá trị nhỏ nhất nhất trong khoảng giá trị muốn chuẩn hóa; x_{min} và x_{max} tương ứng là giá trị nhỏ nhất và giá trị lớn nhất trong n ô tham chiếu, x_j là giá trị muốn chuẩn hóa, y_j là giá trị sau chuẩn hóa x_j ;

bước 3: xây dựng và huấn luyện mô hình học máy phân biệt môi trường; tại bước này, tiến hành huấn luyện mô hình học với hai nhãn lớp là “đồng nhất” và “không đồng nhất”; và sử dụng mô hình này để phân biệt được loại môi trường thu được; cụ thể:

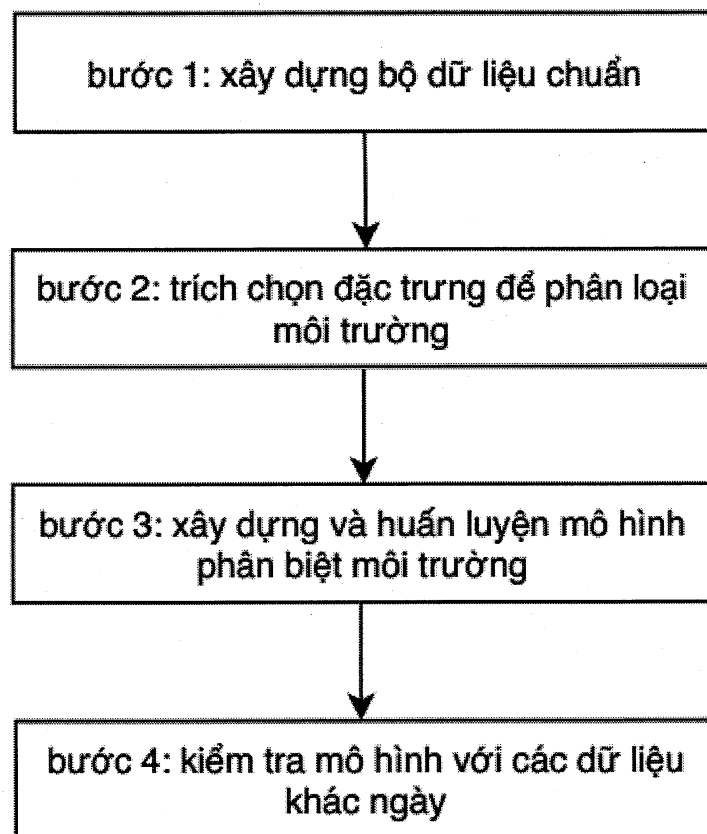
thu một bộ dữ liệu “huấn luyện” với kích thước khoảng một triệu mẫu, trong đó một nửa dữ liệu được gán nhãn “đồng nhất”, một nửa được gán nhãn “không đồng nhất”; quá trình thử nghiệm được thực hiện bằng cách chia bộ dữ liệu thành 2 tập “huấn luyện” và “kiểm thử” với tỉ lệ tương ứng là 80:20, trong đó tập “huấn luyện” sẽ được sử dụng để huấn luyện mô hình học máy và tập “kiểm thử” sẽ được sử dụng để xác định độ chính xác của mô hình;

bước 4: kiểm tra mô hình với các dữ liệu khác ngày; bước này kiểm tra độ chính xác và tin cậy của mô hình đã huấn luyện bằng việc kiểm tra với các bộ dữ liệu khác nhau.

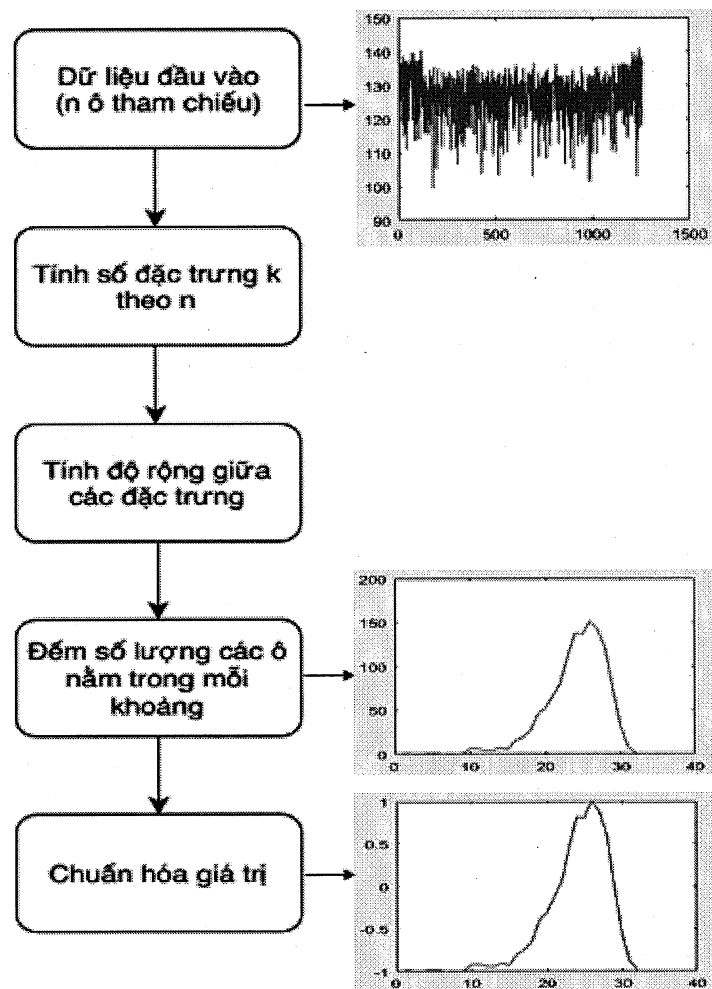
Hình vẽ



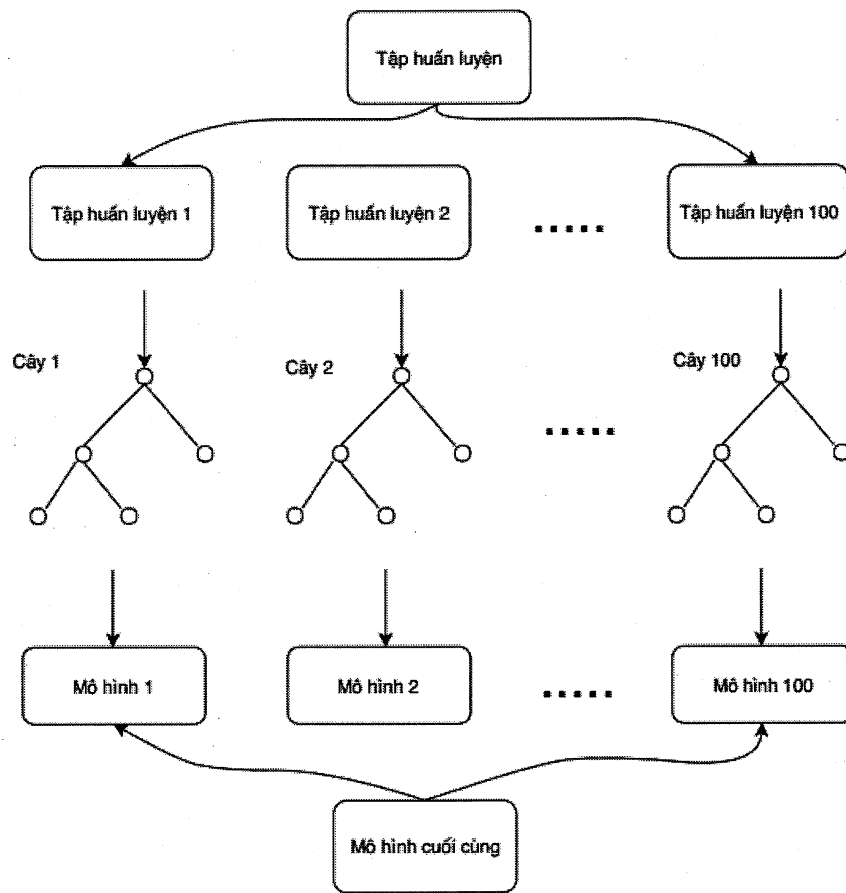
Hình 1



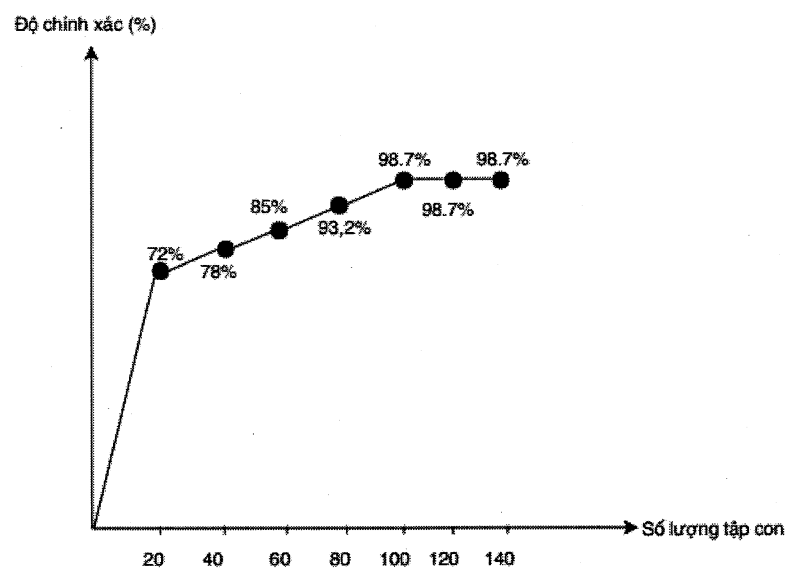
Hình 2



Hình 3



Hình 4



Hình 5