TẠP CHÍ KHOA HỌC TRƯỜNG ĐẠI HỌC BẠC LIÊU
Bac Lieu University Journal of Science
Lĩnh vực Khoa học công nghệ
ISSN 2734-973X | e-ISSN 3093-3420

RỪNG NGẪU NHIÊN XÁC SUẤT VÀ KHẢ NĂNG MÔ HÌNH HÓA SỰ BẤT ĐỊNH TRONG HỌC MÁY

PROBABILISTIC RANDOM FOREST AND ITS ROLE IN MODELING UNCERTAINTY IN MACHINE LEARNING

Huỳnh Phú Sĩ*

Trường Đại học Đồng Tháp
* huynhphusi@gmail.com

Ngày nhận bài: 14/9/2025
Ngày chấp nhận đăng: 23/12/2025

Keywords: *Machine learning, Noisy data, PRF, Probability density function, Probability mass function, Probabilistic Random Forest.*

ABSTRACT

Real-world data often contain uncertainty caused by measurement noise, inconsistent labeling, or incomplete information, which significantly degrades the performance and reliability of traditional machine learning models. To address this limitation, the Probabilistic Random Forest (PRF) has been proposed as a powerful extension of the conventional Random Forest, allowing probability distributions to be directly integrated into the structure of decision trees. PRF models input features as probability density functions (PDFs) and output labels as probability mass functions (PMFs), thereby preserving and utilizing uncertainty information throughout the propagation and classification process - enhancing both predictive accuracy and interpretability. This paper provides a detailed presentation of the mathematical principles underlying PRF, including the probabilistic propagation mechanism at each node, an expected cost function based on an extended Gini impurity for optimal splitting, and a soft voting method for prediction aggregation. The key contribution highlighted in this study is PRF's ability to produce probabilistic output distributions rather than single-label predictions, enabling direct quantification of prediction confidence.

Từ khóa: *Dữ liệu bất định, Hàm khối lượng xác suất, Hàm mật độ xác suất, Học*

TÓM TẮT

Dữ liệu thực tế thường chứa độ bất định do nhiễu đo lường, gán nhãn không nhất quán hoặc thiếu thông tin, làm giảm đáng kể hiệu suất và độ tin cậy của các mô hình học máy truyền thống. Để giải quyết hạn chế này, thuật toán Rừng ngẫu nhiên xác suất (Probabilistic Random Forest – PRF) ra đời như một mở rộng mạnh mẽ của Rừng ngẫu nhiên truyền thống, cho phép tích hợp trực tiếp phân phối xác suất vào kiến trúc cây quyết định. PRF mô hình hóa đặc trưng đầu vào dưới dạng hàm mật độ xác suất (PDF) và nhãn đầu ra dưới dạng hàm khối lượng xác suất (PMF), qua đó duy trì và khai thác thông tin bất định xuyên suốt quá trình lan truyền và phân lớp, từ đó nâng cao độ chính xác và khả năng giải thích. Bài báo này trình bày chi tiết cơ chế hoạt động và các nguyên lý toán học của PRF, bao gồm cơ chế lan truyền xác suất theo từng nút, hàm chi phí kỳ vọng dựa trên Gini impurity mở rộng để tối ưu hóa việc chia nhánh, và phương pháp tổng hợp kết quả dự đoán thông qua bỏ phiếu mềm. Điểm mới quan trọng mà bài báo nhấn mạnh là khả năng PRF cung cấp kết quả dự đoán dưới dạng

máy, PRF, Rừng ngẫu nhiên xác suất.

một phân phối xác suất đầy đủ thay vì một nhãn duy nhất, qua đó cho phép lượng hóa trực tiếp độ tin cậy của dự đoán.

1. Giới thiệu

Trong những năm gần đây, các thuật toán học máy (machine learning – ML) đã trở thành công cụ thiết yếu trong phân tích dữ liệu và hỗ trợ ra quyết định trong nhiều lĩnh vực như y tế, tài chính, thị giác máy tính và an ninh mạng (Chandola, Banerjee, & Kumar, 2009). Một trong những bài toán cốt lõi của ML là phân lớp dữ liệu, trong đó các thuật toán xây dựng mô hình dự đoán nhãn cho các mẫu đầu vào dựa trên đặc trưng đã biết. Một mô hình phân lớp phổ biến và hiệu quả là Rừng ngẫu nhiên (Random Forest – RF), mô hình này được xây dựng dựa trên nguyên lý tổ hợp của nhiều cây quyết định được huấn luyện một cách độc lập (Breiman, 2001; Zhang & Ma, 2012).

Thuật toán RF cho thấy ưu thế về khả năng mở rộng trên dữ liệu mới, khả năng chống quá khớp (overfitting), và độ chính xác cao trong nhiều tình huống thực nghiệm. Tuy nhiên, một hạn chế cơ bản của RF và các mô hình cây truyền thống là giả định dữ liệu đầu vào là những điểm xác định (deterministic), tức là mỗi đặc trưng và nhãn đầu ra đều được coi là giá trị đúng tuyệt đối (Reis & cs., 2018). Trong khi đó, dữ liệu thực tế thường chứa độ bất định (uncertainty) phát sinh từ nhiều nguyên nhân như sai số đo lường, thiếu hụt thông tin hoặc gán nhãn không nhất quán giữa các chuyên gia (Gal, 2016; Kendall & Gal, 2017).

Nhận thức được vấn đề này, nhiều nỗ lực đã được thực hiện nhằm tăng cường khả năng xử lý bất định cho các mô hình cây quyết định, như tích hợp yếu tố xác suất trong voting (Breitenbach & cs., 2003), lựa chọn đặc trưng có trọng số ngẫu nhiên (Gondane & Devi, 2015), hoặc kết hợp RF với các mô hình Bayes (Zhang & Ma, 2012). Tuy nhiên, các cách tiếp cận đó vẫn chưa giải quyết triệt để vấn đề bất định vì vẫn xử lý dữ liệu dưới dạng điểm cố định hoặc mô hình hóa bất định theo cách gián tiếp.

Để lấp đầy khoảng trống này, thuật toán Rừng ngẫu nhiên xác suất (Probabilistic Random Forest – PRF) đã được Reis và cộng sự đề xuất như một giải pháp tích hợp bất định trực tiếp vào kiến trúc của mô hình RF (Reis & cs., 2018). Khác với các biến thể trước đó, PRF biểu diễn mỗi đặc trưng đầu vào như phân phối liên tục (thường là Gaussian), và nhãn đầu ra như phân phối rời rạc. Thay vì phân loại theo kiểu "có hoặc không", mẫu dữ liệu được phân chia thành nhiều khả năng và đồng thời lan truyền theo từng xác suất nhỏ qua các nhánh, tương tự như cách con người đưa ra nhiều phỏng đoán trước một tình huống không chắc chắn.

Bài báo này trình bày và phân tích cơ sở toán học của thuật toán PRF, bao gồm mô hình hóa dữ liệu, chiến lược phân nhánh, hàm chi phí, quá trình suy diễn và cơ chế bỏ phiếu mềm. Qua đó, bài báo làm rõ cách PRF mở rộng kiến trúc RF để xử lý hiệu quả dữ liệu nhiễu, đồng thời cung cấp đầu ra với độ chính xác cao và mức độ tin cậy được lượng hóa rõ ràng – điều đặc biệt quan trọng trong các lĩnh vực như y học, tài chính và tự động hóa.

2. Cơ sở lý thuyết

2.1 Cây quyết định và Rừng ngẫu nhiên truyền thống

Cây quyết định là một mô hình học máy phi tham số, tổ chức dữ liệu theo cấu trúc cây nhị phân, trong đó mỗi nút là một điều kiện kiểm tra trên một đặc trưng, và các nhánh con đại diện cho kết quả của điều kiện đó. Mỗi mẫu dữ liệu được dẫn dắt từ nút gốc đến một nút lá, nơi nó được gán nhãn đầu ra tương ứng. Quá trình xây dựng cây thường sử dụng chiến lược "chia để trị", với tiêu chí lựa chọn đặc trưng và ngưỡng nhằm tối ưu hóa độ thuần khiết (impurity) trong phân bố nhãn tại mỗi nút (Quinlan, 1986). Hai thước đo phổ biến cho độ thuần khiết là Entropy và Gini impurity. Trong bài toán phân lớp, Gini impurity được định nghĩa như sau:

$$G(t) = 1 - \sum_c p_c^2 \qquad (1)$$

trong đó p_c là xác suất (ước lượng theo tần suất) mẫu thuộc lớp c tại nút t, và C là số lớp (Breiman, 1984).

Rừng ngẫu nhiên (RF) là một phương pháp học tổ hợp (ensemble learning) được đề xuất bởi Breiman (2001), mở rộng từ cây quyết định bằng cách kết hợp nhiều cây độc lập để cải thiện hiệu năng dự đoán và tính tổng quát hóa. Thuật toán RF xây dựng một tập hợp gồm T cây quyết định, mỗi cây được huấn luyện trên một tập con bootstrap (tức lấy mẫu ngẫu nhiên có hoàn lại) của dữ liệu gốc, đồng thời sử dụng một tập con ngẫu nhiên của các đặc trưng tại mỗi lần chia nhánh để tăng tính đa dạng (Zhang & Ma, 2012). Trong bài toán phân lớp, đầu ra của RF được tính bằng hình thức bỏ phiếu đa số (majority voting):

$$\hat{y} = \arg\max_{y \in \mathrm{Y}} \sum_{t} \mathrm{I}\left[h_t(x) = y \right] \qquad (2)$$

trong đó $h_t(x)$ là kết quả phân lớp của cây thứ t, Y là tập nhãn, và I[·] là hàm chỉ báo, trả về 1 nếu điều kiện đúng và 0 nếu sai.

RF thừa hưởng ưu điểm từ cây quyết định như khả năng xử lý cả dữ liệu rời rạc và liên tục, đồng thời khắc phục nhược điểm dễ overfitting nhờ vào tính ngẫu nhiên và tổ hợp. Tuy nhiên, RF vẫn giữ nguyên giả định rằng đặc trưng và nhãn trong tập dữ liệu là giá trị xác định. Điều này dẫn đến hạn chế khi áp dụng RF trong các bài toán có nhiễu đo lường hoặc nhãn không chắc chắn – một đặc điểm phổ biến trong các dữ liệu thực tế (Reis & cs., 2018; Gal, 2016).

2.2. Vấn đề bất định trong học máy

Trong hầu hết các mô hình học máy truyền thống, dữ liệu huấn luyện thường được giả định là hoàn toàn xác định, tức mỗi giá trị đặc trưng và nhãn được xem là chính xác tuyệt đối. Tuy nhiên, trong thực tế, dữ liệu hiếm khi hoàn hảo. Các hệ thống đo lường, quy trình thu thập hoặc gán nhãn thường bị ảnh hưởng bởi nhiễu, sai số, thiếu hụt hoặc tính chủ quan, tạo ra những yếu tố bất định không thể bỏ qua (Gal, 2016; Kendall & Gal, 2017).

Bất định trong học máy là biểu hiện của mức độ thiếu chắc chắn về giá trị của đặc trưng đầu vào, nhãn đầu ra hoặc bản thân mô hình dự đoán. Tùy theo nguyên nhân phát sinh, bất định được chia thành hai loại chính (Gal, 2016):

- Bất định nội tại (aleatoric): Xuất phát từ bản chất ngẫu nhiên không thể loại bỏ của dữ liệu, ví dụ như nhiễu đo lường từ cảm biến hoặc nhãn không nhất quán do con người gán nhãn khác nhau cho cùng một đối tượng.
- Bất định mô hình (epistemic): Xuất phát từ thiếu hụt dữ liệu hoặc kiến thức mô hình, có thể cải thiện được bằng cách thu thập thêm dữ liệu hoặc cải thiện cấu trúc mô hình.

Khi bất định không được mô hình hóa đúng cách, hiệu suất và độ tin cậy của mô hình học máy có thể bị suy giảm. Bất định trong dữ liệu đầu vào hoặc nhãn, như trong ảnh y học, có thể dẫn đến suy diễn sai và giảm độ chính xác. Ngoài ra, các mô hình không xử lý được bất định sẽ thiếu khả năng ước lượng độ tin cậy trong dự đoán, gây rủi ro nghiêm trọng trong các ứng dụng như y tế, tài chính hoặc lái xe tự hành.

3. Cơ sở toán học của thuật toán PRF

3.1 Mô hình hóa dữ liệu bằng phân phối xác suất

Một điểm đổi mới cốt lõi trong thuật toán Rừng ngẫu nhiên xác suất (PRF) là khả năng mô hình hóa trực tiếp độ bất định trong dữ liệu, thông qua việc biểu diễn đặc trưng và nhãn dưới dạng các phân phối xác suất. Điều này cho phép PRF tiếp nhận và khai thác thông tin không chắc chắn một cách tự nhiên ngay trong quá trình học và suy diễn, thay vì chỉ điều chỉnh kết quả đầu ra như một số mô hình trước đó (Reis & cs., 2018).

3.1.1 Biểu diễn đặc trưng đầu vào bằng phân phối liên tục (PDF)

Trong PRF, mỗi đặc trưng x_{ij} của mẫu dữ liệu thứ i được giả định là một biến ngẫu nhiên liên tục tuân theo phân phối chuẩn (Gaussian):

$$X_{ij} \sim \mathrm{N}\,(\mu_{ij}, \sigma_{ij}^2) \qquad (3)$$

Trong đó, μ_{ij} là giá trị trung bình đặc trưng thứ j của mẫu i, σ_{ij}^2 là phương sai thể hiện mức độ không chắc chắn hoặc sai số đo lường của đặc trưng thứ j ở mẫu i.

Chẳng hạn, một đặc trưng nhiệt độ được đo là $37.0^{\circ}C$ với sai số thiết bị $\pm 0.2^{\circ}C$ có thể được mô hình hóa dưới dạng $x \sim \mathrm{N}\,(37.0, 0.2^2)$. Việc sử dụng phân phối Gaussian giúp đơn giản hóa quá trình suy luận và lan truyền xác suất trong cây, nhờ các tính chất giải tích của hàm mật độ xác suất chuẩn. Tại mỗi nút i, PRF không thực hiện phân nhánh cứng như cây truyền thống, mà tính xác suất để mẫu rơi vào từng nhánh dựa trên phân phối của đặc trưng. Điều này giống như việc đánh giá khả năng một điểm nằm bên trái ($X_{ij} \le \theta$) hay bên phải ($X_{ij} > \theta$) ngưỡng, thay vì khẳng định chắc chắn.

$$\pi_i^{(L)} = P\left(X_{ij} \le \theta\right) = \Phi\left(\frac{\theta - x_{ij}}{\sigma_{ij}}\right) \quad \text{và} \quad \pi_i^{(R)} = P(X_{ij} > \theta) = 1 - \pi_i^{(L)}$$

Với $\Phi(\cdot)$ là hàm phân phối tích lũy (CDF) của phân phối chuẩn tắc, còn $\pi_i^{(L)}, \pi_i^{(R)}$ lần lượt là xác suất để mẫu lan truyền sang nhánh trái/phải với ngưỡng chia θ. Mỗi mẫu do đó không đi theo một đường duy nhất mà lan tỏa song song qua nhiều nhánh, phản ánh sự không chắc chắn trong đặc trưng đầu vào. Ví dụ, giả sử một đặc trưng x_{ij} có giá trị đo được là 5.0 với độ lệch chuẩn $\sigma_{ij} = 1.0,$ và ngưỡng chia tại nút là $\theta = 6.0,$ khi đó:

$$\pi_i^{(L)} = \Phi\left(\frac{6.0 - 5.0}{1.0}\right) = \Phi(1.0) \approx 0.8413 \quad \text{và} \quad \pi_i^{(R)} = 1 - \pi_i^{(L)} \approx 0.1587$$

Theo đó, hay vì kết luận đối tượng này truyền sang nhánh trái, PRF cho rằng đối tượng có 84.13% khả năng được lan truyền sang nhánh trái và 15.87% khả năng sang nhánh phải.

3.1.2 Biểu diễn nhãn đầu ra bằng phân phối rời rạc (PMF)

Về phía nhãn huấn luyện, PRF không yêu cầu mỗi mẫu phải được gán một nhãn duy nhất, mà cho phép biểu diễn bằng hàm khối lượng xác suất (PMF). Cụ thể, với tập nhãn $\mathrm{Y} = \{C_1, C_2, .., C_K\}$, mỗi mẫu i có thể có phân phối nhãn:

$$\mathrm{PMF}_i = \left\{ P(y_i = C_k) = p_{i,k} \middle| \sum_k p_{i,k} = 1 \right\} \qquad (5)$$

Trong đó, $p_{i,k}$ là xác suất mà mẫu i thuộc về lớp C_k. Chẳng hạn, thay vì gán nhãn cứng "dương tính" hay "âm tính" cho một ca chẩn đoán y tế không rõ ràng, người ta có thể gán phân phối nhãn cho ca này dưới dạng:

$$\mathrm{PMF}: P(y = \text{positive}) = 0.6,\ P(y = \text{negative}) = 0.4$$

Cách biểu diễn này rất phù hợp với dữ liệu có nhãn bị nhiễu, dữ liệu bán giám sát hoặc dữ liệu được gán nhãn bởi nhiều chuyên gia với độ tin cậy khác nhau (Kendall & Gal, 2017).

3.2 Quá trình lan truyền theo xác suất

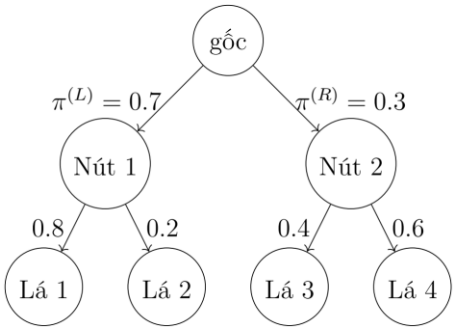
Trong cây quyết định truyền thống, mỗi mẫu được phân nhánh một cách tất định tại mỗi nút dựa trên điều kiện dạng $x_j \le \theta.$ Tuy nhiên, trong môi trường dữ liệu bất định, đặc trưng đầu vào X_j được xem là một biến ngẫu nhiên thay vì một giá trị đơn lẻ. Do đó, PRF áp dụng một chiến lược chia nhánh theo xác suất dựa trên phân phối xác suất của đặc trưng.

Giả sử một đặc trưng X_j tuân theo phân phối chuẩn $X_j \sim \mathrm{N}\,(\mu_j, \sigma_j^2)$. Tại một nút phân chia n sử dụng ngưỡng θ_n, xác suất để mẫu đi sang nhánh trái (tức là $X_j \le \theta_n$) được xác định bởi

hàm phân phối tích lũy $\pi_n^{(L)}$ và $\pi_n^{(R)}$. Khi đó, xác suất để một mẫu lan truyền đến một nút cụ thể n trong cây được tính bằng tích các xác suất rẽ nhánh trên đường đi từ gốc đến n:

$$\pi_n = \prod_{m\in \mathrm{L}} \pi_m^{(L)} \cdot \prod_{m\in \mathrm{R}} \pi_m^{(R)} = \prod_{m\in \mathrm{L}} \Phi\left(\frac{\theta_m - \mu_{j_m}}{\sigma_{j_m}}\right) \cdot \prod_{m\in \mathrm{R}} \left[1 - \Phi\left(\frac{\theta_m - \mu_{j_m}}{\sigma_{j_m}}\right)\right]$$

Trong đó, R , L là tập các nút đi sang trái hoặc phải trên đường từ nút gốc đến nút n; θ_m, j_m là ngưỡng chia và chỉ số đặc trưng tại nút m. Hình 1 minh họa quá trình lan truyền của một mẫu đến các nút lá:



Hình 1. Minh họa lan truyền xác suất của một mẫu trong cây PRF

Trong Hình 1 ta thấy, xác suất để một mẫu lan truyền đến các nút lá là:

$\pi_{Lá 1} = \pi_{\text{gốc}}^{(L)} \cdot \pi_{\text{Nút 1}}^{(L)} = 0.56$; $\pi_{Lá 2} = \pi_{\text{gốc}}^{(L)} \cdot \pi_{\text{Nút 1}}^{(R)} = 0.14$;

$\pi_{Lá 3} = \pi_{\text{gốc}}^{(R)} \cdot \pi_{\text{Nút 2}}^{(L)} = 0.12$; $\pi_{Lá 4} = \pi_{\text{gốc}}^{(R)} \cdot \pi_{\text{Nút 2}}^{(R)} = 0.18$

3.3 Hàm chi phí và tiêu chí chia nhánh

Để xây dựng cây, PRF cần chọn ra đặc trưng và ngưỡng phân chia sao cho hai nhánh sau khi chia càng "thuần khiết" càng tốt. Chỉ số đánh giá thuần khiết trong PRF là một phiên bản mở rộng của Gini impurity. Giả sử tại một nút n, mỗi mẫu i mang xác suất lan truyền $\pi_n^{(i)}$, và nhãn của mẫu là một phân phối rời rạc $\{p_{i,c}\}_{c=1}^{C}$, trong đó C là số lớp. Khi đó, phân phối nhãn trung bình tại nút n được tính như sau:

$$\bar{p}_{n,c} = \frac{1}{Z_n} \sum_i \pi_n^{(i)} \cdot p_{i,c}, \text{với } Z_n = \sum_i \pi_n^{(i)} \quad (5)$$

Với phân phối nhãn trung bình như trên, Gini impurity tại nút n được xác định bởi:

$$G(n) = 1 - \sum_c \bar{p}_{n,c}^2 \quad (6)$$

Khi chia nút thành hai nhánh trái (L) và phải (R), hàm chi phí tổng dựa trên đặc trưng j và ngưỡng θ được tính như sau:

$$\mathrm{C}(j,\theta) = \frac{Z_L}{Z_P} G(L) + \frac{Z_R}{Z_P} G(R) \quad (7)$$

Trong đó, Z_L, Z_R là tổng xác suất lan truyền các mẫu vào nhánh trái/phải; $Z_P = Z_L + Z_R$ là tổng xác suất tại nút hiện tại (nút cha – trước khi phân chia); $G(L)$, $G(R)$ là Gini impurity của hai nhánh. Xét một ví dụ đơn giản gồm 3 mẫu với 1 đặc trưng X, có các thông tin sau:

Bảng 1. Dữ liệu giả lập minh họa quy trình chọn ngưỡng phân chia

Mẫu	μ	σ	Nhãn
1	3.0	0.5	$[A:1.0, B:0.0]$
2	4.0	0.5	$[A:0.0, B:1.0]$
3	5.0	0.5	$[A:0.5, B:0.5]$

Xét một ngưỡng phân chia $\theta = 4.2$. Tại nút gốc:

Mẫu 1: $\pi_1^{(L)} = \Phi\left(\frac{4.2-3.0}{0.5}\right) \approx 0.9918$, $\pi_1^{(R)} = 1-0.9918 \approx 0.0082.$

Mẫu 2: $\pi_2^{(L)} = \Phi\left(\frac{4.2-4.0}{0.5}\right) \approx 0.6554$, $\pi_2^{(R)} = 1-0.6554 \approx 0.3446.$

Mẫu 3: $\pi_3^{(L)} = \Phi\left(\frac{4.2-5.0}{0.5}\right) \approx 0.0548$, $\pi_3^{(R)} = 1-0.0548 \approx 0.9452.$

Khi đó:

$$Z_L = \sum_i \pi_i^{(L)} = 0.9918 + 0.6554 + 0.0548 = 1.702;$$

$$Z_R = \sum_i \pi_i^{(R)} = 0.0082 + 0.3446 + 0.9452 = 1.298;$$

$$Z_P = Z_L + Z_R = 3.000$$

Bằng cách gộp tất cả các PMF theo trọng số lan truyền vào nhánh trái, ta tính được phân phối nhãn trung bình tại nhánh trái như sau:

$$\bar{p}_{L,A} = \frac{0.9918 \cdot 1.0 + 0.6554 \cdot 0.0 + 0.0548 \cdot 0.5}{1.702} \approx 0.5987;$$

$$\bar{p}_{L,B} = 1 - \bar{p}_{L,A} = 1 - 0.5987 = 0.4013.$$

Vậy Gini tại nhánh trái là $G(L) = 1 - (0.5987^2 + 0.4013^2) \approx 0.4805.$

Tương tự, với nhánh phải ta có:

$$\bar{p}_{R,A} = \frac{0.0082 \cdot 1.0 + 0.3446 \cdot 0.0 + 0.9452 \cdot 0.5}{1.298} \approx 0.3704;$$

$$\bar{p}_{R,B} = 1 - \bar{p}_{R,A} = 1 - 0.3704 \approx 0.6296.$$

Suy ra $G(R) = 1 - (0.3704^2 + 0.6296^2) \approx 0.4664.$

Khi đó, chi phí tổng với ngưỡng $\theta = 4.2$ là

$$\mathrm{C}(j, 4.2) = \frac{1.702}{3.0} \cdot 0.4805 + \frac{1.298}{3.0} \cdot 0.4664 \approx 0.4744.$$

Ta có thể lặp lại các bước trên cho nhiều giá trị θ khác nhau, rồi chọn ngưỡng có C nhỏ nhất làm điểm phân chia tối ưu. Như vậy, PRF thực hiện tìm kiếm ngưỡng phân chia bằng cách đánh giá toàn bộ ngưỡng khả dĩ theo giá trị chi phí, thay vì chọn dựa trên điểm cắt dữ liệu như cây quyết định truyền thống. Kỹ thuật này cho phép PRF khai thác trọn vẹn thông tin xác suất và đưa ra quyết định chia nhánh ổn định hơn so với các mô hình cây truyền thống, đặc biệt khi dữ liệu có nhiễu.

3.4 Suy diễn và tổng hợp kết quả

Với mô hình PRF $\mathrm{F} = \{t_1, t_2, \ldots, t_T\}$ gồm T cây xác suất đã huấn luyện, mục tiêu là dự đoán phân phối nhãn cho mẫu mới x_{test}, trong đó mỗi đặc trưng cũng được biểu diễn dưới dạng phân phối xác suất, ví dụ $x_j \sim \mathrm{N}(\mu_j, \sigma_j^2)$. Quá trình suy diễn gồm hai giai đoạn:

3.4.1 Giai đoạn 1: Suy diễn trên từng cây xác suất

Đối với mỗi cây $t \in \mathrm{F}$, mục tiêu là tính toán phân phối xác suất hậu nghiệm $\hat{P}_t(y \mid x_{\text{test}})$. Quá trình này thực hiện như sau:

a) Lan truyền xác suất tới các nút lá: Xác suất để x_{test} đến được một nút lá ℓ cụ thể được tính bằng tích của các xác suất rẽ nhánh trên đường đi từ gốc đến lá ℓ:

$$\pi_\ell(x_{\text{test}}) = \prod_{n\in\text{path}(\ell)} \pi_n(x_{\text{test}}) = \prod_{n\in\text{L}} \pi_n^{(L)}(x_{\text{test}}) \cdot \prod_{n\in\text{R}} \pi_n^{(R)}(x_{\text{test}}) \quad (8)$$

trong đó path(ℓ) là tập hợp các nút trên đường đi đến lá ℓ, $\pi_n(x_{\text{test}})$ là xác suất để x_{test} đi theo nhánh tương ứng tại nút n, R và L là tập các nút đi sang trái hoặc phải. Cần lưu ý rằng tổng xác suất đến tất cả các lá của một cây luôn bằng 1.

b) Tổng hợp phân phối tại mỗi cây: Phân phối dự đoán của toàn bộ cây t cho mẫu x_{test} được tính bằng trung bình có trọng số của các phân phối tại các nút lá, với trọng số chính là xác suất lan truyền $\pi_\ell(x_{test})$:

$$\hat{P}_t(y \mid x_{\text{test}}) = \sum_{\ell\in L^{\wedge}(t)} \pi_\ell(x_{\text{test}}) \cdot P_\ell(y) \quad (9)$$

Trong đó, $Lá(t)$ là tập các nút lá trong cây thứ t, $P_\ell(y)$ là phân phối xác suất tại nút lá ℓ của nhãn y. Kết quả của giai đoạn này là một tập hợp gồm T phân phối xác suất, mỗi phân phối tương ứng với một cây trong rừng.

3.4.2 Giai đoạn 2: Tổng hợp kết quả toàn cục

Trong PRF, việc tổng hợp các phân phối dự đoán từ từng cây trong rừng để tạo thành phân phối đầu ra duy nhất cho mẫu kiểm tra x_{test} được thực hiện thông qua cơ chế voting mềm, tức là tính trung bình cộng của các phân phối xác suất từ tất cả T cây:

$$\hat{P}_{\text{final}}(y \mid x_{\text{test}}) = \frac{1}{T}\sum_t \hat{P}_t(y \mid x_{\text{test}}) \quad (10)$$

Phân phối $\hat{P}_{\text{final}}(y \mid x_{\text{test}})$ chính là kết quả đầu ra cuối cùng của mô hình PRF. Nó không chỉ chứa thông tin về lớp có khả năng xảy ra cao nhất mà còn phản ánh mức độ không chắc chắn trong dự đoán, tạo cơ sở cho việc đánh giá độ tin cậy của dự đoán.

4. Thảo luận

Với RF truyền thống, đặc trưng và nhãn được xem là các điểm dữ liệu xác định, trong khi đó, thuật toán PRF mô hình hóa mỗi đặc trưng đầu vào X_{ij} dưới dạng biến ngẫu nhiên liên tục theo phân phối chuẩn $\text{N}(\mu_{ij}, \sigma_{ij}^2)$, và nhãn y_i dưới dạng phân phối rời rạc. Cách tiếp cận này cho phép PRF phản ánh tốt hơn tính chất không hoàn hảo của dữ liệu thực tế, bao gồm nhiễu đo lường, sai lệch nhãn và thông tin không đầy đủ. Thay vì phân nhánh quyết định tại mỗi nút như RF, PRF lan truyền mỗi mẫu đồng thời đến cả hai nhánh, với xác suất được tính từ hàm phân phối tích lũy (CDF) của đặc trưng tại ngưỡng chia. Cơ chế lan truyền xác suất này giúp bảo toàn thông tin bất định trong suốt quá trình học, thay vì làm mất nó như trong các cây quyết định truyền thống.

Ngoài ra, thay vì chỉ đưa ra một nhãn phân lớp duy nhất, PRF tổng hợp phân phối xác suất từ tất cả các nút lá mà mẫu có thể đi tới, từ đó tạo ra kết quả là phân phối $P(y \mid x)$. Điều này không chỉ giúp mô hình ước lượng được mức độ tin cậy cho từng dự đoán, mà còn cải thiện khả năng xử lý trong các tình huống phân lớp khó hoặc nhãn mơ hồ.

Dù có nền tảng lý thuyết vững chắc, PRF cũng tồn tại một số hạn chế đáng lưu ý. Thứ nhất, việc tính toán xác suất lan truyền, phân phối nhãn tại mỗi nút và tổng hợp đầu ra bằng kỳ vọng (trung bình có trọng số theo xác suất) khiến cho chi phí tính toán của PRF cao hơn đáng kể so với RF truyền thống. Chi phí này tăng theo số lượng mẫu, số đặc trưng và độ sâu của cây. Hơn nữa, để PRF hoạt động hiệu quả, mô hình cần được cung cấp thông tin về mức độ bất định trong giá trị đầu vào, cụ thể là phương sai hoặc độ lệch chuẩn của từng đặc trưng. Đây là một giả định không phải lúc nào cũng khả thi trong thực tế, đặc biệt với các bộ dữ liệu không được thu thập một cách có kiểm soát.

5. Kết luận

Bài viết đã phân tích các nguyên lý và cơ chế toán học của thuật toán Rừng ngẫu nhiên xác suất, một mở rộng của Rừng ngẫu nhiên truyền thống, thuật toán này cho phép mô hình hóa trực tiếp độ bất định trong dữ liệu thông qua phân phối xác suất. Việc lan truyền mẫu theo xác suất, sử dụng hàm chi phí dựa trên kỳ vọng, và kết luận bằng phân phối mềm giúp PRF xử lý hiệu quả dữ liệu nhiễu và nhãn không chắc chắn.

Tuy nhiên, bên cạnh các ưu điểm lý thuyết, PRF cũng đối mặt với thách thức về chi phí tính toán và yêu cầu thông tin thống kê đầu vào. Trong tương lai, việc tối ưu hóa thuật toán và tích hợp vào các hệ thống học máy hiện đại sẽ là hướng nghiên cứu tiềm năng nhằm phát huy tối đa khả năng của mô hình này trong thực tế.

TÀI LIỆU THAM KHẢO

Breiman, L. (1984). *Classification and regression trees*. CRC press.

Breiman, L. (2001). *Random Forests.* Machine Learning 45 (1), 5–32.

Breitenbach, Markus & Nielsen, Rodney & Grudic, Gregory. (2003). *Probabilistic Random Forests: Predicting Data Point Specific Misclassification Probabilities*. University of Colorado at Boulder.

Chandola, Varun & Banerjee, Arindam & Kumar, Vipin. (2009). *Anomaly Detection: A Survey*. ACM Comput. Surv.. 41.

Gal, Y. (2016). *Uncertainty in deep learning* (Doctoral dissertation, University of Cambridge).

Gondane, R., & Devi, V. S. (2015). *Classification using probabilistic random forest*. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 174–179). IEEE.

Kendall, A., & Gal, Y. (2017). *What uncertainties do we need in Bayesian deep learning for computer vision?* In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 5580–5590). Curran Associates Inc.

Quinlan, J. R. (1986). *Induction of decision trees*. Machine Learning, 1(1), 81–106.

Reis, Itamar & Baron, Dalya & Shahaf, Sahar. (2018). *Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets*. The Astronomical Journal. 157. 16.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications*. Springer.