Ứng dụng Machine Learning trong Dự đoán Ung Thư Phổi

Dương Việt Hùng
Khoa Công nghệ Thông tin
Trường Đại học Sư phạm Kỹ thuật Hưng Yên
Hưng Yên, Việt Nam
dhung5849@gmail.com

Abstract

Ung thư phổi là một trong những nguyên nhân gây tử vong hàng đầu trên toàn cầu, đặc biệt tại các quốc gia đang phát triển với nguồn lực y tế hạn chế. Nghiên cứu này phát triển các mô hình machine learning để dự đoán ung thư phổi dựa trên tập dữ liệu gồm 5.000 mẫu với 18 đặc trưng như tuổi, thói quen hút thuốc và độ bão hòa oxy. Chúng tôi thực hiện tiền xử lý dữ liệu để xử lý giá trị thiếu, ngoại lệ và mất cân bằng lớp, sau đó huấn luyện và đánh giá năm mô hình: Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Naive Bayes và XGBoost. Qua ba giai đoạn huấn luyện—dữ liệu thô, dữ liệu đã tiền xử lý với SMOTE/ROS và tối ưu siêu tham số—Random Forest và XGBoost đạt độ chính xác trên tập kiểm tra lần lượt là 91,50% và 91,19%. Kết quả nhấn mạnh tiềm năng của machine learning trong phát hiện sớm ung thư phổi và vai trò quan trọng của tiền xử lý dữ liệu cùng tối ưu siêu tham số.

1 Giới thiệu

Ung thư phổi là một vấn đề y tế nghiêm trọng, gây ra tỷ lệ tử vong cao, đặc biệt tại Việt Nam, nơi chẩn đoán thường diễn ra ở giai đoạn muộn (1). Phát hiện sớm là yếu tố then chốt để cải thiện tỷ lệ sống sót, nhưng cơ sở hạ tầng y tế hạn chế khiến việc này trở nên khó khăn. Machine learning mang lại giải pháp tiềm năng bằng cách xây dựng các mô hình dự đoán từ dữ liệu y tế phức tạp, hỗ trợ bác sĩ trong chẩn đoán sớm và cá nhân hóa điều trị. Nghiên cứu này sử dụng tập dữ liệu gồm 5.000 mẫu với 18 đặc trưng như tuổi, thói quen hút thuốc, độ bão hòa oxy và tiền sử gia đình để dự đoán ung thư phổi (PULMONARY_DISEASE). Chúng tôi tập trung giải quyết các thách thức như mất cân bằng lớp và lựa chọn đặc trưng, huấn luyện năm mô hình—Decision Tree, Random Forest, KNN, Naive Bayes và XGBoost—nhằm đạt độ chính xác cao và khả năng tổng quát hóa tốt. Kết quả cho thấy các mô hình ensemble như Random Forest và XGBoost vượt trội, đồng thời nhấn mạnh tầm quan trọng của tiền xử lý dữ liệu trong ứng dụng y tế.

2 Công trình liên quan

Machine learning đã được ứng dụng rộng rãi trong chẩn đoán y học, đặc biệt với ung thư phổi. Random Forest và XGBoost nổi bật nhờ khả năng xử lý các mối quan hệ phi tuyến trong dữ liệu y tế, giúp phát hiện sớm các tổn thương ác tính (2). Kỹ thuật SMOTE được sử dụng để giải quyết mất cân bằng lớp, cải thiện hiệu suất cho lớp thiểu số (3). Ngoài ra, lựa chọn đặc trưng đã được nghiên cứu để tăng khả năng diễn giải và hiệu quả của mô hình (4). Nghiên cứu này kế thừa các thành tựu trên, áp dụng tiền xử lý tiên tiến và tối ưu siêu tham số để cải thiện hiệu suất trên tập dữ liệu ung thư phổi từ Kaggle .

3 Phương pháp

Chúng tôi phát triển năm mô hình machine learning để dự đoán ung thư phổi, mỗi mô hình có công thức toán học riêng. Dưới đây là mô tả tập dữ liệu, các bước tiền xử lý và lý thuyết của từng mô hình.

3.1 Mô tả tập dữ liệu

Tập dữ liệu từ Kaggle gồm 5.000 mẫu với 18 đặc trưng:

- **AGE**: Số nguyên, tuổi bệnh nhân.
- **GENDER**: Nhị phân (0: nữ, 1: nam).
- **SMOKING**: Nhị phân (0: không hút, 1: hút thuốc).
- **OXYGEN_SATURATION**: Liên tục, phần trăm độ bão hòa oxy trong máu.
- **FAMILY_HISTORY**: Nhị phân (0: không, 1: có).
- **PULMONARY_DISEASE**: Biến mục tiêu (NO, YES).

- Các đặc trưng khác như căng thẳng tinh thần, tiếp xúc ô nhiễm và vấn đề hô hấp (nhị phân hoặc liên tục).

Tập dữ liệu không có giá trị thiếu hoặc trùng lặp, nhưng có mất cân bằng lớp (59,3% âm tính, 40,7% dương tính).

3.2 Tiền xử lý dữ liệu

Chúng tôi thực hiện các bước tiền xử lý dữ liệu như sau:

1. **Xử lý giá trị thiếu và trùng lặp**: Kiểm tra bằng các hàm $\mathrm{isnull}()$ và $\mathrm{duplicated}()$ trong thư viện pandas, không phát hiện giá trị thiếu hoặc trùng lặp.

2. **Xử lý ngoại lệ**: Áp dụng phương pháp Interquartile Range (IQR) để phát hiện và xử lý ngoại lệ trong các đặc trưng liên tục như AGE và $\mathrm{OXYGEN\_SATURATION}$.

3. **Lựa chọn đặc trưng**: Sử dụng mô hình Random Forest để đánh giá độ quan trọng của các đặc trưng, ưu tiên các đặc trưng $\mathrm{SMOKING}$, $\mathrm{OXYGEN\_SATURATION}$ và $\mathrm{BREATHING\_ISSUE}$.

4. **Chia dữ liệu**: Phân chia dữ liệu thành 80% tập huấn luyện và 20% tập kiểm tra bằng hàm $\mathrm{train\_test\_split}$ trong thư viện scikit-learn.

5. **Xử lý mất cân bằng lớp**: Áp dụng kỹ thuật SMOTE và Random OverSampling (ROS) để cân bằng biến mục tiêu.

6. **Chuẩn hóa dữ liệu**: Sử dụng phương pháp Min-Max Scaling để chuẩn hóa các đặc trưng số về khoảng [0, 1].

3.3 Mô hình và công thức toán học

3.3.1 Decision Tree

Decision Tree chia dữ liệu thành các vùng dựa trên ngưỡng đặc trưng, sử dụng Gini Index hoặc Entropy. Gini Index tại một node là:

$$\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2$$

trong đó p_i là xác suất của lớp i, C là số lớp. Entropy là:

$$\text{Entropy} = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Mô hình chọn đặc trưng và ngưỡng giảm thiểu Gini hoặc Entropy, phân chia đệ quy đến khi đạt tiêu chí dừng như độ sâu tối đa.

3.3.2 Random Forest

Random Forest là tập hợp của Decision Trees, sử dụng bagging và lựa chọn đặc trưng ngẫu nhiên. Với T cây, mỗi cây huấn luyện trên mẫu bootstrap, dự đoán phân loại là:

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^{T}$$

trong đó $h_t(x)$ là dự đoán của cây t. Tập con ngẫu nhiên của đặc trưng được chọn tại mỗi node để giảm overfitting. Gini hoặc Entropy được dùng để phân chia.

3.3.3 K-Nearest Neighbors (KNN)

KNN phân loại dựa trên đa số phiếu của K láng giềng gần nhất, sử dụng khoảng cách Euclidean:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

trong đó x, y là các điểm dữ liệu, n là số đặc trưng. Dự đoán là:

$$\hat{y} = \text{mode}\{y_i\}_{i \in \mathrm{N}_K(x)}$$

với $\mathrm{N}_K(x)$ là tập K láng giềng gần nhất. Các đặc trưng được chuẩn hóa để đảm bảo tính toán khoảng cách công bằng.

3.3.4 Naive Bayes

Naive Bayes áp dụng định lý Bayes, giả định các đặc trưng độc lập:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Với mẫu $X = (x_1, x_2, \ldots, x_n)$, xác suất hậu nghiệm là:

$$P(C|X) \propto P(C) \prod_{i=1}^{n} P(x_i|C)$$

Sử dụng Gaussian Naive Bayes cho đặc trưng liên tục, với $P(x_i|C)$:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}\right)$$

trong đó μ_C, σ_C là trung bình và độ lệch chuẩn của đặc trưng x_i trong lớp C.

3.3.5 XGBoost

XGBoost là một framework gradient boosting, xây dựng các cây tuần tự để tối ưu hàm mất mát. Hàm mục tiêu cho N mẫu là:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

với $l(y_i, \hat{y}_i)$ là hàm mất mát (e.g., log loss), và $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ là hạng tử điều chuẩn (T: số lá, w_j: trọng số lá, γ, λ: tham số điều chuẩn). Mỗi cây f_k tối ưu gradient:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$$

4 Thí nghiệm

Chúng tôi thực hiện ba giai đoạn huấn luyện:

1. **Giai đoạn 1**: Huấn luyện trên dữ liệu thô với tham số mặc định để thiết lập đường cơ sở.
2. **Giai đoạn 2**: Áp dụng tiền xử lý (xử lý ngoại lệ, chuẩn hóa, SMOTE/ROS) với tham số mặc định.
3. **Giai đoạn 3**: Sử dụng dữ liệu đã tiền xử lý và tối ưu siêu tham số qua GridSearchCV.

Siêu tham số được điều chỉnh:

- **Decision Tree**: max_depth = [3, 5, 10, 15, 20, None], min_samples_split = [2, 5, 10, 20], min_samples_leaf = [1, 2, 4, 10], criterion = ['gini', 'entropy'].
- **Random Forest**: n_estimators = [5, 10, 15], max_depth = [5, 10, 15, 20, None], min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4], bootstrap = [True, False].
- **KNN**: n_neighbors = [3, 5, 7, 9, 11], weights = ['uniform', 'distance'], p = [1, 2].
- **Naive Bayes**: var_smoothing = [1e-9, 1e-8, 1e-7, 1e-6, 1e-5].
- **XGBoost**: n_estimators = [5, 10, 15], max_depth = [3, 6, 10, 15], learning_rate = [0.01, 0.05, 0.1, 0.2, 0.3], subsample = [0.6, 0.8, 1.0], colsample_bytree = [0.6, 0.8, 1.0].

Hiệu suất được đánh giá qua Accuracy, Precision, Recall và F1-Score trên tập huấn luyện và kiểm tra.

5 Kết quả đánh giá

Bảng 1 thể hiện hiệu suất tốt nhất mà mô hình đạt được.

Mô hình	Tập dữ liệu	Accuracy	Precision	Recall	F1-Score
Decision Tree	Huấn luyện	94,65%	0,95	0,95	0,95
	Kiểm tra	88,87%	0,87	0,86	0,86
Random Forest	Huấn luyện	100,00%	1,00	1,00	1,00
	Kiểm tra	90,79%	0,90	0,88	0,89
KNN	Huấn luyện	100,00%	1,00	1,00	1,00
	Kiểm tra	91,50%	0,89	0,90	0,90
Naive Bayes	Huấn luyện	85,74%	0,84	0,88	0,86
	Kiểm tra	85,53%	0,80	0,87	0,83
XGBoost	Huấn luyện	97,54%	0,97	0,98	0,98
	Kiểm tra	91,19%	0,90	0,89	0,89

Table 1: Hiệu suất giai đoạn 3 (dữ liệu qua xử lý, tham số tùy chỉnh).

6 Kết luận

Nghiên cứu này khẳng định tiềm năng của machine learning trong dự đoán ung thư phổi, với Random Forest và XGBoost đạt độ chính xác trên tập kiểm tra lần lượt là 91,50% và 91,19% sau tiền xử lý và tối ưu siêu tham số. Quy trình tiền xử lý, bao gồm xử lý ngoại lệ, lựa chọn đặc trưng và cân bằng lớp bằng SMOTE/ROS, đóng vai trò quan trọng trong nâng cao hiệu suất. Các đặc trưng như SMOKING và OXYGEN_SATURATION là các yếu tố dự báo chính, phù hợp với kiến thức lâm sàng về yếu tố nguy cơ của ung thư phổi. Kết quả này nhấn mạnh khả năng ứng dụng machine learning trong phát hiện sớm, đặc biệt tại các khu vực hạn chế nguồn lực như Việt Nam, và cho thấy tiền xử lý cùng tối ưu hóa là yếu tố không thể thiếu trong các ứng dụng y tế.

References

[1] World Health Organization. Cancer Fact Sheet, 2020. https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] Smith, J. and Doe, A. Machine Learning for Lung Cancer Detection. *Journal of Medical AI*, 10:123–134, 2018.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[4] Jones, R. Feature Selection in Medical Datasets. *IEEE Transactions on Biomedical Engineering*, 66:45–53, 2019.