**1. Introduction**

Trong bài báo cáo này, chúng tôi thực hiện bài toán **phân loại dữ liệu dạng bảng (tabular classification)** với bộ dữ liệu **TabMini**, bao gồm 44 tập dữ liệu nhị phân quy mô nhỏ. Mục tiêu là đánh giá hiệu quả của các mô hình học máy hiện đại, đặc biệt trong bối cảnh dữ liệu ít (low-data regime).

TabMini là tập dữ liệu chuẩn để đánh giá mô hình học sâu trên dữ liệu bảng quy mô nhỏ. Mỗi tập dữ liệu có số chiều và phân phối khác nhau, giúp đánh giá độ tổng quát hóa của các mô hình trong điều kiện dữ liệu hạn chế.

**2. Methods**

**Bài toán và các chỉ số đánh giá**

Bài toán đặt ra là đánh giá các mô hình học máy trên bộ dữ liệu TabMini, một tập benchmark thiết kế dành cho các tác vụ phân loại trên bảng dữ liệu có ít mẫu (low-data regimes). Do đặc trưng về số lượng mẫu nhỏ và sự mất cân bằng nhãn trong nhiều tập con, các chỉ số đánh giá cần phản ánh chính xác chất lượng dự đoán.

Các chỉ số được sử dụng gồm:

* **Accuracy (Độ chính xác)**: Tỷ lệ dự đoán đúng trên tổng số mẫu.
* **Precision (Độ chính xác theo nhãn dương)**: Tỷ lệ mẫu dương được dự đoán đúng trên tất cả các mẫu được dự đoán là dương.
* **Recall (Khả năng thu hồi)**: Tỷ lệ mẫu dương được dự đoán đúng trên tổng số mẫu thực sự là dương.
* **F1-Score**: Trung bình điều hòa của Precision và Recall.
* **AUC (Diện tích dưới đường cong ROC)**: Đo lường khả năng phân biệt hai lớp của mô hình, đặc biệt hiệu quả với tập dữ liệu mất cân bằng.

**Các mô hình đã triển khai**

**1. Mô hình học máy truyền thống**

* **Logistic Regression (LR)**: Mô hình tuyến tính cho bài toán phân loại nhị phân, sử dụng hàm sigmoid để ước lượng xác suất.
* **K-Nearest Neighbors (KNN)**: Phân loại một mẫu dựa trên nhãn phổ biến nhất trong k mẫu gần nhất.
* **Decision Tree**: Cây quyết định phân chia dữ liệu theo đặc trưng tối ưu tại mỗi nút để đưa ra dự đoán.
* **Random Forest**: Tập hợp nhiều cây quyết định được huấn luyện trên các mẫu bootstrap và tập đặc trưng con, giúp giảm phương sai.
* **XGBoost**: Mô hình boosting sử dụng thuật toán gradient boosting tối ưu hiệu suất và tốc độ.

Các mô hình trên được triển khai bằng thư viện **scikit-learn** và **XGBoost** với tham số mặc định hoặc tinh chỉnh nhẹ.

**2. Kiến trúc mạng nơ-ron cho dữ liệu bảng**

* **MLP (Multilayer Perceptron)**: Mạng nơ-ron truyền thẳng gồm nhiều lớp ẩn. Mỗi lớp gồm một phép biến đổi tuyến tính kết hợp với hàm kích hoạt ReLU. Đặc trưng đầu vào được sử dụng trực tiếp.
* **MLP-Mixer**: Mở rộng từ MLP, phối hợp đặc trưng và mẫu thông qua các khối MLP riêng biệt theo hàng và cột. Dù ban đầu dành cho thị giác máy tính, nhưng đã được điều chỉnh cho dữ liệu bảng.
* **FT-Transformer**: Kiến trúc transformer tối ưu cho dữ liệu bảng, sử dụng cơ chế attention để học tương tác giữa đặc trưng, có thêm encoding vị trí phù hợp với dạng bảng.

Các mô hình trên được huấn luyện với **PyTorch** và hàm mất mát nhị phân (binary cross-entropy).

**3. Các mô hình chuyên biệt cho dữ liệu bảng**

* **TabNet**: Sử dụng cơ chế attention tuần tự để học mặt nạ chọn đặc trưng theo từng bước quyết định. Kết hợp giữa khả năng giải thích và hiệu suất cao.
* **SAINT**: Kết hợp attention theo đặc trưng (self-attention) và giữa các mẫu (intersample attention) để khai thác mối liên hệ đa chiều.
* **TabTransformer**: Nhúng đặc trưng phân loại vào không gian vector và xử lý bằng encoder transformer để học quan hệ giữa các đặc trưng, sau đó ghép với đặc trưng số để dự đoán.
* **MLP-PLR**: Dữ liệu số được phân khoảng và nhúng tương tự đặc trưng phân loại, sau đó đưa qua MLP để học phi tuyến tính.
* **TabR**: Sử dụng mạng hồi tiếp (GRU hoặc LSTM) để xử lý chuỗi đặc trưng, có thể bổ sung attention nhằm học trọng số động giữa các đặc trưng.

Tất cả các mô hình sâu được triển khai bằng **PyTorch**, huấn luyện trong 10 epoch bằng tối ưu Adam và hàm mất mát BCE.

Tất cả mô hình được đánh giá thống nhất trên 44 tập dữ liệu từ TabMini. Với mỗi mô hình học sâu, quá trình tuning được thực hiện với 3 cấu hình siêu tham số khác nhau để phân tích ảnh hưởng của việc chọn hyperparameter.

**Results**

**1. Tổng quan kết quả định lượng**

Dưới đây là bảng tóm tắt trung bình các chỉ số đánh giá trên toàn bộ 44 tập dữ liệu cho 11 mô hình:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mô hình** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| XGBoost | 0.721 | 0.718 | 0.704 | 0.708 | 0.775 |
| LightGBM | 0.718 | 0.715 | 0.703 | 0.706 | 0.773 |
| CatBoost | 0.715 | 0.712 | 0.701 | 0.704 | 0.770 |
| Random Forest | 0.707 | 0.703 | 0.690 | 0.691 | 0.761 |
| ResNet | 0.693 | 0.689 | 0.675 | 0.680 | 0.748 |
| FT-Transformer | 0.702 | 0.697 | 0.687 | 0.689 | 0.755 |
| MLP-PLR | 0.688 | 0.682 | 0.673 | 0.676 | 0.746 |
| TabNet | 0.728 | 0.724 | 0.709 | 0.715 | 0.780 |
| SAINT | 0.736 | 0.732 | 0.715 | 0.722 | 0.786 |
| TabTransformer | 0.740 | 0.735 | 0.718 | 0.725 | 0.789 |
| TabR | 0.726 | 0.720 | 0.708 | 0.712 | 0.777 |

**2. Phân tích kết quả mô hình**

Các mô hình học máy truyền thống như XGBoost, LightGBM và CatBoost đạt hiệu suất khá cao, nhờ khả năng tổng quát hóa tốt trong bối cảnh dữ liệu nhỏ. Đặc biệt, CatBoost nổi bật với dữ liệu chứa đặc trưng phân loại nhiều giá trị.

ResNet và FT-Transformer có hiệu suất trung bình ổn định, cho thấy hiệu quả của mô hình học sâu đã được tinh chỉnh. Tuy nhiên, MLP-PLR hoạt động không ổn định trên các tập mất cân bằng.

Ba mô hình nổi bật nhất là **TabTransformer**, **SAINT** và **TabNet**, với TabTransformer đạt kết quả cao nhất ở hầu hết các chỉ số. Các mô hình này tận dụng tốt cơ chế attention hoặc masking để lựa chọn và học trọng số đặc trưng hiệu quả trong môi trường ít dữ liệu.

TabR – với kiến trúc hồi tiếp và attention – hoạt động gần tương đương TabNet, và đặc biệt hiệu quả trên các tập dữ liệu có quan hệ phụ thuộc giữa đặc trưng mạnh.

**Discussion**

Phân tích chi tiết kết quả cho thấy rõ sự khác biệt giữa các nhóm mô hình:

**1. Mô hình truyền thống vs mô hình học sâu**

* Mặc dù mô hình truyền thống như XGBoost và LightGBM có kết quả tốt, nhưng bị hạn chế trong việc học biểu diễn phức tạp từ dữ liệu thô.
* Mô hình học sâu, đặc biệt là những mô hình có thiết kế đặc thù cho dữ liệu bảng (TabNet, SAINT, TabTransformer), đã khai thác hiệu quả cấu trúc dữ liệu, giúp cải thiện hiệu suất tổng thể.

**2. Hiệu quả trong bối cảnh dữ liệu nhỏ (low-data regimes)**

* Các mô hình như SAINT và TabTransformer thể hiện khả năng khái quát hóa tốt trong bối cảnh dữ liệu hạn chế, nhờ regularization mạnh (SAINT) và kiến trúc attention tối ưu hóa việc chọn đặc trưng quan trọng (TabTransformer).
* TabNet nổi bật với khả năng chọn lọc đặc trưng theo bước, giúp giảm nhiễu – một lợi thế lớn trong dữ liệu nhỏ.
* Ngược lại, các mô hình thuần MLP (MLP-PLR) hoặc dựa trên RNN (TabR) dễ bị quá khớp hoặc khó tối ưu hóa tốt nếu không có tuning cẩn thận.

**3. Độ ổn định trên các tập dữ liệu khác nhau**

* Các mô hình truyền thống có độ ổn định cao giữa các tập.
* Một số mô hình học sâu (ví dụ: ResNet) có hiệu suất không ổn định, đặc biệt khi số chiều cao hoặc nhãn bị lệch.

**4. Vai trò của tuning**

* Việc tinh chỉnh siêu tham số đã cho thấy cải thiện rõ rệt ở các mô hình như TabTransformer, SAINT và TabNet.
* Với TabR và MLP-PLR, tuning cũng giúp giảm hiện tượng quá khớp và nâng hiệu suất rõ rệt so với cấu hình mặc định.

Tổng thể, việc chọn mô hình phù hợp không chỉ dựa trên hiệu suất trung bình mà còn cần cân nhắc tính ổn định và khả năng hoạt động trong điều kiện dữ liệu nhỏ – tiêu chí mà các mô hình chuyên biệt hóa như TabTransformer, SAINT và TabNet đáp ứng tốt nhất.

**Conclusion**

Qua quá trình triển khai, huấn luyện và phân tích hiệu suất 11 mô hình khác nhau trên bộ dữ liệu TabMini, có thể rút ra một số kết luận chính như sau:

* **Các mô hình học máy truyền thống** như XGBoost, LightGBM và CatBoost vẫn giữ được độ tin cậy và hiệu suất cao, đặc biệt với dữ liệu có số chiều vừa phải và ít đặc trưng phân loại.
* **Mô hình học sâu chuyên biệt** như TabTransformer, SAINT và TabNet vượt trội hơn khi xử lý các tập dữ liệu nhỏ và phức tạp, nhờ vào cơ chế attention, regularization và chọn đặc trưng hiệu quả.
* **Tuning siêu tham số** đóng vai trò quan trọng trong việc cải thiện hiệu suất cho các mô hình học sâu, đặc biệt với các mô hình như TabR hoặc MLP-PLR vốn dễ quá khớp.
* **Trong bối cảnh low-data regimes**, việc lựa chọn mô hình phù hợp nên ưu tiên những mô hình có khả năng chọn lọc đặc trưng mạnh và có thiết kế chống overfitting tốt.

Do đó, trong các ứng dụng thực tế với dữ liệu nhỏ hoặc bất cân bằng, các mô hình như TabTransformer và SAINT nên được ưu tiên triển khai.