ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Dự đoán bệnh tan máu bằng mô hình học máy và học sâu

Phạm Anh Minh
minh.pa194802@sis.hust.edu.vn

Ngành: Công nghệ thông tin

Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang _____

Chữ kí GVHD

Khoa: Kỹ thuật máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 06/2024

LỜI CẢM ƠN

Trong suốt quá trình nghiên cứu và thực hiện đồ án, em đã nhận được rất nhiều sự hướng dẫn, giúp đỡ tận tình của thầy cô, anh chị cùng các bạn. Với lòng biết ơn sâu sắc, em xin được gửi lời cảm ơn đến Ban giám hiệu Đại học Bách Khoa Hà Nội nói chung, ban giám hiệu Trường Công nghệ Thông tin và Truyền thông nói riêng đã tạo một môi trường học tập và đào tạo tích cực, chuyên nghiệp nhưng vẫn rất thân thiện. Em xin gửi lời cảm ơn đến toàn bộ đội ngũ giáo viên giảng dạy của trường nói chung và giảng viên trường Công nghệ Thông tin và Truyền thông nói riêng đã giúp em có được cơ sở lý thuyết vững chắc và tự tin hơn trong nghề nghiệp của bản thân. Đặc biệt, em xin gửi lời cảm ơn chân thành đến thầy Nguyễn Hồng Quang, bộ môn Kỹ thuật máy tính, trường Công nghệ Thông tin và Truyền thông, người đã trực tiếp hướng dẫn, tận tình chỉ bảo và đưa ra những lời khuyên vô cùng hữu ích trong suốt quá trình làm đồ án giúp em có thể hoàn thành đồ án một cách tốt nhất.Xin cảm ơn bố mẹ, anh chị và bạn bè đã luôn là nguồn động viên to lớn để con yên tâm học tập và đến hôm nay đã hoàn thành đồ án tốt nghiệp. Cuối cùng, xin gửi lời cảm ơn đến chính bản thân trong 5 năm học đã cố gắng nỗ lực không ngừng. Một lần nữa, em xin chân thành cảm ơn!

TÓM TẮT NỘI DUNG ĐỒ ÁN

Sepsis is a severe complication arising from an infection. If not treated promptly, it can result in organ failure and death. Therefore, early detection and treatment of sepsis can potentially save many lives. However, their effectiveness often depends on the awareness and acceptance of these procedures. In this study, i implement a sepsis check based on the widely accepted Sepsis-3 guidelines. My implementation achieved an F-score of up to 0.874. Alongside the rule-based approach for early sepsis detection, i also employ the existing data-driven transformer-based STraTS model (Tipirneni and Reddy, 2021) for time-series forecasting to support sepsis checks and directly predict sepsis using 24-hour patient data in a fully data-driven setup. Furthermore, i aim to enhance the mono-modal STraTS model by incorporating a clinical text embedding module to enable multi-modal learning. Both the original STraTS model and my refined STraTS+Text model performed well in forecasting (with a masked MSE of approximately 5.24) and classification tasks (with a ROC-AUC of approximately 0.89).

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

ABSTRACT

Mục này khuyến khích sinh viên viết lại mục “Tóm tắt” đồ án tốt nghiệp ở trang trước bằng tiếng Anh. Phần này phải có đầy đủ các nội dung như trong phần tóm tắt bằng tiếng Việt. Sinh viên không nhất thiết phải trình bày mục này.

Nhưng nếu lựa chọn trình bày, sinh viên cần đảm bảo câu từ và ngữ pháp chuẩn tắc, nếu không sẽ có tác dụng ngược, gây phản cảm.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
API	Giao diện lập trình ứng dụng (Application Programming Interface)
BTS	Trạm thu phát sóng di động(Base Transceiver Station)
GPS	Hệ thống Định vị Toàn cầu (Global Positioning System)
HTTP	Giao thức truyền tải siêu văn bản (HyperText Transfer Protocol)
LOS	đường truyền tần số vô tuyến không che khuất bởi các chướng ngại vật (Line-of-Sight)
MS	Thiết bị di động(Mobile station)
NLOS	đường truyền tần số vô tuyến bị che khuất bởi các chướng ngại vật (Non-Line-of-Sight)

CHAPTER 1. INTRODUCTION

Sepsis occurs when the body's immune response to an infection becomes dysregulated, leading to systemic inflammation. It is a leading cause of death in Intensive Care Units (ICU). Early detection is crucial for patient survival (Rudd et al., 2020). To identify septic patients from their clinical data, Singer et al. (2016) and Reyna et al. (2019) present slightly different rule-based guidelines focusing on suspected infections and clinical criteria for life-threatening organ dysfunction. Singer et al.'s (2016) guidelines were developed as an in-hospital tool to assess patient condition.These guidelines can be applied to observed data and to forecast time-series values, potentially allowing earlier identification and prevention of sepsis. I implement a rule-based sepsis check based on the widely accepted Sepsis-3 guidelines to enable early sepsis prediction.

Additionally, I refine an existing Self-supervised Transformer for Time-Series (STraTS) model (Tipirneni and Reddy, 2021) for time-series forecasting and 24-hour sepsis prediction. The STraTS regression model forecasts time-series values following each observation window to support the sepsis check, while the STraTS classification model predicts sepsis using 24-hour patient data in a fully data-driven setup. I enhance the original STraTS architecture, which only takes continuous physiological features as input, by integrating a clinical text embedding module based on Clinical BERT (Alsentzer et al., 2019) to encode 1.4 million clinical notes from patients in my MIMIC-IV data.Both models (STraTS and STraTS+Text) perform well in forecasting and classification tasks, achieving a masked Mean Squared Error (MSE) of approximately 5.24 and a ROC-AUC of approximately 0.89.

My rule-based sepsis check achieved an F-score of up to 0.874 without using features predicted by the STraTS forecasting model. While introducing predicted values from the STraTS forecasting model did not improve the rule-based sepsis check's performance, these predictions helped address data sparsity, enabling the rule-based check to identify septic patients whose clinical data alone would not have been sufficient for accurate classification.