Bước tiến lớn trong phân tích dữ liệu y tế điện tử bằng phương pháp xử lý ngôn ngữ tự nhiên và học sâu

Sự tiến bộ trong phân tích dữ liệu y tế điện tử bằng phương pháp xử lý ngôn ngữ tự nhiên và học sâu.

1. Giới thiệu.
   Bài báo này sẽ đưa ra giải pháp để xử lý dữ liệu chưa có cấu trúc, phát hiện các thuật ngữ chuyên ngành bằng các mô hình không gian vector, so sánh mô hình học sâu với các nhiệm vụ y tế và sử dụng mô hình học sâu để nhận diện vật thể, hoặc thiết lập mối liên hệ giữa các đặc tính và các lớp tương ứng của các ảnh.
2. Xử lý ngôn ngữ tự nhiên (NLP) cho dữ liệu y tế điện tử.
   Phân tích văn bản bao gồm sử dụng các thuật toán nhằm hiểu nội dung văn bản. Trong y tế điện tử, giải pháp NLP thường sử dụng cho những tác vụ như: sàng lọc, chẩn đoán điều trị, giám sát bệnh nhân… bằng cách áp dụng vào các mô hình y tế điện tử phức tạp như hệ thống chẩn đoán bằng máy tính, hệ thống hỗ trợ quyết định hướng dữ liệu.
   Xử lý NLP gồm 3 bước chính: tiền xử lý, mô hình không gian vector và triển khai mô hình.

   2.1 Công nghệ mới nhất của tiền xử lý văn bản y tế điện tử.

   Công nghệ giải mã viết tắt là phương pháp giải mã những từ viết tắt thành cụm từ hoàn chỉnh. Có 3 cách tiếp cận phổ biến: thuật toán bỏ phiếu trong nhận diện thuật ngữ, thuật toán học không giám sát nhận diện thuật ngữ, thuật toán dựa trên từ ngữ nhúng.

   Đánh giá n-gram sử dụng mô hình xác suất để dự đoán phần tiếp theo của n-gram, trong đó n-gram là chuỗi phần tử liên tiếp của một phần văn bản. Ý tưởng là sử dụng thống kê tần suất chữ cái để phát hiện chuỗi ký tự bất thường.

   Giải pháp mới nhất cho sửa lỗi chính tả theo ngữ cảnh là sử dụng mô hình ngôn ngữ nhúng để học sự tương quan giữa lỗi chính tả và từ chính xác tương ứng.

   2.2 Mô hình không gian vector hiệu quả cho dữ liệu y tế.

   Huấn luyện mô hình thông qua việc thay đổi các tham số của nó để phù hợp với các văn bản huấn luyện. Tìm số vòng lặp để mô hình không bị "quá tương thích" rất quan trọng. Thời gian huấn luyện mô hình kéo dài 4-5 tiếng, có thể làm giảm thời gian huấn luyện bằng cách sử dụng các mô hình đã huấn luyện sẵn và huấn luyện với dữ liệu tự cung cấp. Tuy nhiên điều này có thể làm giảm độ chính xác.

   Những mô hình phổ biến và công cụ để trích xuất đặc tính sử dụng từ nhúng: Word2Vec Embedding, FastText Embedding, BERT Embedding, Sense-Disambiguation Embedding.

   2.3 Đánh giá kiến trúc học sâu trong mảng y tế.

   Một số mô hình mạng thần kinh học sâu phổ biến là được dùng trong giải pháp y tế điện tử: mạng thần kinh tích chập (CNN), mạng thần kinh hồi quy (RNN), bộ nhớ dài-ngắn hạn (LSTM).

   CNN được dùng chủ yếu trong xử lý ảnh, được dùng để giải quyết các vấn đề y tế, vd phát hiện hành vi bất thường của con người… RNN được dùng để chuyển lời nói thành văn bản, dịch máy và mô hình hóa ngôn ngữ. LSTM tương đồng với RNN, nhưng lưu trữ dữ liệu cũ trong mô hình dễ dàng hơn.

   Kiểm chứng chéo là phương pháp thống kê nhằm đo lường khả năng khái quát của mô hình trên một tập dữ liệu độc lập. Mô hình được xây dựng trên tập huấn luyện và kiểm thử trên tập kiểm thử.

   Đánh giá mô hình giúp đánh giá độ chính xác của mô hình. Kết quả đánh giá nên được xem xét chuyên gia có hiểu biết và chi tiết về trường hợp đang quan sát. Sau khi được huấn luyện và đánh giá, mô hình cần vượt qua tiêu chuẩn vàng. Nếu chuyên gia quan sát được mô hình không hoạt động chính xác, mô hình sẽ được chỉnh sửa và huấn luyện lại thông qua các phản hồi của chuyên gia.

3. Kỹ thuật học sâu cho xử lý ảnh y tế.
   Để trích xuất dữ liệu có ích từ ảnh, ngoài các mô hình học sâu đã nói ở 2.3, có thể dùng một số kỹ thuật khác như kĩ thuật đặt tag.
   Các nhiệm vụ xử lý ảnh có thể phân loại thành hai loại chính: phân chia và phân loại. Ngoài ra, kiến trúc mạng thần kinh còn được dùng cho mục tiêu khác như tiền xử lý ảnh. Đa số CNN và LSTM được dùng để xử lý ảnh y tế.
4. Kết luận.
   Kho dữ liệu khổng lồ từ các bệnh viện có thể dùng để cải thiện dịch vụ y tế, tuy nhiên phần lớn các dữ liệu đều chưa qua xử lý và các dữ liệu cá nhân phải được giữ bí mật. Trong phân tích dữ liệu y tế, sự phức tạp còn đến từ tính không đồng nhất của dữ liệu. May mắn, những phương pháp mới đang liên tục được đưa ra nhằm trích xuất thông tin, từ văn bản y tế đến ảnh.