Tóm tắt

Bài toán xử lý dữ liệu khuyết là một trong những bài toán đáng quan tâm trong thực tế. Dữ liệu khuyết thường xảy ra trong quá trình thu thập dữ liệu (data collection), đặc biệt với các dữ liệu chuỗi thời gian, thu thập từ các ứng dụng thứ ba, và các dữ liệu từ các cảm biến (sensors) của các thiết bị IoT. Việc xử lý dữ liệu khuyết hiệu quả cũng góp phần cải thiện chất lượng dữ liệu đầu vào và góp phần cải tiến chất lượng của các thuật toán Máy học được đề xuất. Chính vì lý do đó, bài toán xử lý dữ liệu khuyết cũng thu hút nhiều quan tâm của các nhà nghiên cứu trong lĩnh vực thống kê, khoa học dữ liệu, máy học ứng dụng, và khai phá dữ liệu.

Với mức độ phổ biến của vấn đề dữ liệu khuyết đã có rất nhiều công cụ và gói lệnh trong python được phát triển để đáp ứng nhu cầu của người dùng, có thể kể đến như: gói lệnh fancyimpute (sử dụng các kỹ thuật như Nuclear Norm Minimization, SOFT-IMPUTE, Iterative SVD để xử lý dữ liệu), gói lệnh missingpy (sử dụng phương pháp missForest để điền vào dữ liệu khuyết), gói lệnh autoimpute (chứa các hàm cho phép người dùng khảo sát các kiểu khuyết xuất hiện trong bảng dữ liệu và chứa nhiều phương pháp cho cả dữ liệu khuyết nhiều chiều, một chiều và chuỗi thời gian), gói lệnh gcimpute (dựa trên mô hình Gaussian copula xử lý dữ liệu có dạng dòng chảy (streaming data) bị khuyết),... Tuy nhiên, dù có nhiều nghiên cứu mới được ra đời nhưng trong đó lại có rất nhiều nghiên cứu không công khai code, không xuất bản gói lệnh, hoặc không cung cấp chi tiết cách sử dụng code được cung cấp. Điều này gây khó khăn cho các nhà khoa học không nghiên cứu sâu về dữ liệu khuyết.

Trong đề tài này, chúng tôi tập trung nghiên cứu các thuật toán xử lý dữ liệu khuyết trong các dạng dữ liệu có cấu trúc và phân bố khác nhau, gồm có dữ liệu mất cân bằng (imbalanced data), dữ liệu chuỗi thời gian (time-series data), và dữ liệu đa phương thức (multimodal data). Để thực hiện mục tiêu đó, chúng tôi lên kế hoạch giải quyết các bài toán gồm có 05 mục tiêu chính sau đây:

1. Xây dựng cơ sở dữ liệu cho các bài toán xử lý dữ liệu khuyết.

2. Nghiên cứu và cải tiến thuật toán ước lượng tham số khi có dữ liệu khuyết.

3. Nghiên cứu và áp dụng phương pháp xử lý dữ liệu khuyết trong trường hợp dữ liệu mất cân bằng.

4. Nghiên cứu và áp dụng phương pháp xử lý dữ liệu khuyết cho dữ liệu dạng chuỗi thời gian (time-series data).

5. Nghiên cứu và giải quyết vấn đề dữ liệu khuyết đối với dữ liệu đa phương thức (multimodal).