**MÔN HỌC PHÂN TÍCH CHUỖI THỜI GIAN**

**PHIẾU TRẢ LỜI CÂU HỎI**

Tên đề tài: Một số Phương pháp Sinh dữ liệu cho bài toán Phân lớp chuỗi thời gian

Nhóm: **03**

STT Thuyết trình: **01**

|  |
| --- |
| Câu hỏi (Thầy Đạt): Nhóm đánh giá rằng, không có một phương pháp nào là chuẩn. Liệu có cách nào để đưa ra được rằng, với bộ dữ liệu có dạng như thế nào thì chúng ta nên dùng phương pháp gì không? |
| Trả lời:   * Đầu tiên, kết luận quan trọng nhất của bài báo cáo là “Các phương pháp sinh dữ liệu đem lại hiệu quả khi sử dụng trong huấn luyện các mô hình”. * Tiếp theo, để đưa ra kết luận một phương pháp nào đó là tốt nhất mà chỉ sử đánh giá trên một bộ dữ liệu nhỏ, không có tính khái quát và với một số lượng nhỏ các mô hình là không khách quan. Do vậy, Nhóm quyết định không đưa ra một kết luận rằng “Phương pháp A là tốt nhất”. Tuy vậy, đối với bộ dữ liệu này và với các mô hình cụ thể ta hoàn toàn có thể đưa ra một khẳng định nhỏ như vậy (ví như mô hình *VGG* với phương pháp *Jittering* cho kết quả tốt hơn cả). * Cuối cùng, để có thể đưa ra một nhận định rằng “với bộ dữ liệu có dạng A thì ta nên dùng phương pháp X” – đây là một công việc vô cùng khó đối với Nhóm. Bởi khi này, Nhóm cần thực hiện thu thập một lượng rất lớn dữ liệu, rồi từ đó phân loại các dạng của phương pháp và thực nghiệm từng phương pháp này với đa dạng các mô hình dự đoán. Do thời gian hạn hẹp, cũng như chi phí có thể bỏ cho bài báo cáo này là không quá lớn nên Nhóm chưa thể thực hiện được điều này.   Tuy nhiên, đây cũng là một gợi ý rất hay, rất hữu ích cho Nhóm, để Nhóm tiếp tục phát triển đề tài này trong thời gian tới.   * Ngoài ra, đối với một số phương pháp đơn giản, ta có thể dễ dàng thấy được phạm vi áp dụng của chúng, ví như: phương pháp *rotation, Jittering* thường được áp dụng cho bài toán có đầu vào *dạng ảnh*,... |

|  |
| --- |
| Câu hỏi (Hoàng Phi Long): Quy tắc lấy mẫu huấn luyện ít như vậy, dù tập test có kết quả thấp thì cũng có thể do tập huấn luyện không đại diễn được tập thể chứ không phải do kỹ thuật sinh không tốt? |
| Trả lời:   * Đầu tiên, chúng ta không thể khẳng định được rằng tập huấn luyện có ít mẫu hay nhiều mẫu thì tập đó sẽ đại diện cho tập thể.   Để minh họa, ta có thể đưa ra nhận định rằng: nếu ta thu thập 500 bộ dữ liệu đa dạng về thói quen học tập của sinh viên trên cả nước sẽ có ý nghĩa hơn rất nhiều so với việc chúng ta thu thập thói quen học tập của 5000 sinh viên trường Đại học Bách khoa Hà Nội.   * Ngoài ra, bài báo cáo của Nhóm nhằm đánh giá các phương pháp sinh dữ liệu chứ không đánh giá chất lượng của dữ liệu. Nếu như dữ liệu không khái quát được cho tổng thể thì kết quả của các mô hình khi sử dụng tập dữ liệu gốc và tập dữ liệu mới “sinh” sẽ giảm, nhưng vẫn đảm bảo được kết quả khi sử dụng các tập mới “sinh” tốt hơn kết quả khi sử dụng tập gốc. |

|  |
| --- |
| Câu hỏi (Cô Ngọc Anh): Các phương pháp này sinh dữ liệu này sinh được chuỗi thời gian, vậy có sinh được nhãn không? |
| Trả lời:   * Để trả lời cho câu hỏi này, Nhóm xin mô tả lại quy trình thực nghiệm của bài báo cáo:  |  | | --- | | + 30 bản ghi trong tập huấn luyện, mỗi bản ghi gồm 2 phần {nhãn; chuỗi thời gian}  + Tách Tập huấn luyện thành 2 phần: X{30 chuỗi thời gian}, Y {30 nhãn tương ứng}  + Phương pháp sinh Dữ liệu: input: X, ouput: X’ {3000 bản ghi}  Thực hiện lặp lại 1 phương pháp sinh 99 lần  + Tập Y’ thu được bằng cách: sao chép tập Y 99 lần và ghép vào 1 mảng Y’ đảm bảo thứ tự  + Từ đó, ta thu được tập X’, Y’ để đưa vào mô hình huấn luyện. |  * Do vậy, Nhóm xin trả lời câu hỏi này là phương pháp **có** sinh nhãn. |

|  |
| --- |
| Câu hỏi (Cô Ngọc Anh): Tiêu chí nào để chứng minh nhãn của mình là đúng ? |
| Trả lời:   * Các phương pháp sinh dữ liệu này Nhóm nghiên cứu, tham khảo dựa trên paper dưới đây. Trong bài báo, tác giả không để cập đến việc chứng minh tính đúng đắn của bộ dữ liệu được sinh ra. * Về việc chứng minh tính đúng đắn của các phương pháp, Nhóm chưa tìm ra cách chứng minh bằng cơ sở toán học. * Tuy nhiên, về quan điểm của Nhóm, sau khi tìm hiểu thì Nhóm thấy rằng các phương pháp sinh ra chuỗi thời gian mới này không làm thay đổi bản chất của chuỗi thời gian gốc. Do vậy mà nhãn của các bộ dữ liệu về cơ bản được đảm bảo.   Ví dụ như phương pháp xoay, việc ta xoay ảnh của một con mèo một góc x thì bản chất đó vẫn là con mèo mà khó có thể nhầm sang một con vật khác được (nếu sai xót thì tỉ lệ là chấp nhận được).   * Paper Nhóm đề xuất cô và các bạn đọc nếu muốn hiểu sâu hơn về các phương pháp này: *Iwana, Brian Kenji, and Seiichi Uchida*. "An empirical survey of data augmentation for time series classification with neural networks." Plos one 16.7 (2021): e0254841. |

|  |
| --- |
| Câu hỏi (Thầy Đạt): *DGW* phức tạp hơn *RGW* có đúng không? *Reference* có rất nhiều cái, Nhóm chọn như thế nào? |
| Trả lời:   * DGW phức tạp hơn *RGW* bởi *DGW* có một bước cải tiến ở phần *Reference.* * Về cách chọn *Reference*, Nhóm đã trình bày trong phần báo cáo và xin phép được trích lại: |

|  |
| --- |
| Câu hỏi (Thầy Đạt): Tại sao nhóm chỉ lấy 30 bộ làm training và 900 bộ làm testing mà không làm ngược lại? |
| Trả lời:   * Đầu tiên, nếu ta dùng ngược lại, tức 900 bộ huấn luyện và 30 bộ kiểm định thì dữ liệu huấn luyện của chúng ta đã quá đủ với bộ kiểm định nhỏ như vậy. * Thứ hai, việc thu thập một tập dữ liệu huấn luyện nhỏ như vậy là hoàn toàn khả thi nếu không muốn nói là tương đối dễ. Và đây cũng chính là tính ứng dụng thực tế của bài báo cáo. * Thứ ba, mục đích của bài báo cáo là kiểm tra xem các phương pháp có đạt hiệu quả hay không, tức dữ liệu sinh ra phải đáp ứng được yêu cầu của dữ liệu tổng quan để khi mô hình dự đoán ta thu được kết quả đúng.   Do vậy, đối với bài báo cáo này nhóm quyết định chia bộ huấn luyện nhỏ như vậy mà không lấy khoảng 70% tập dữ liệu như thông thường. |