THƯ TRẢ LỜI PHẢN BIỆN BÀI BÁO 358-1293-1-SP

Kính gửi : Ban biên tập Chuyên san các công trình nghiên cứu và triển khai công nghệ thông tin và truyền thông cùng các phản biện.

Chúng tôi xin trân trọng cảm ơn ban biên tập và các phản biện đã góp ý cho nội dung bài báo 358-1293-1-SP. Chúng tôi đã chỉnh sửa lại nội dung bài báo theo góp ý của các phản biện và xin được giải trình một số điểm như sau:

I. Giải trình góp ý của phản biện 1

1. *Tác giả cần cho biết với bộ dữ liệu thực về phim, cụ thể, có bao nhiêu đặc trưng sản phẩm và đặc trưng người dùng*?

Bộ dữ liệu thực về phim 6040 người dùng cho 3952 phim. Số đặc trưng nội dung thể loại phim được cung cấp là 19, 46 nước sản xuất phim trong CSDL, 72 hãng phim, 734 đạo diễn. Đối với thông tin người dùng, chúng tôi sử dụng 4 đặc trưng về độ tuổi (tuổi<=15, 15<tuổi<=35; 35<tuổi<=60 và tuổi>60), 10 đặc trưng nghề nghiệp và giới tính. Tất cả các đặc trưng này đã được cung cấp đầy đủ trong tập MovieLens. Đây là những đặc trưng nội dung phim và người dùng được tách ra trong CSDL để thử nghiệm. Những nội dung này chúng tôi đã bổ sung thêm vào nội dung bài báo.

2. *Đề xuất của bài báo cần đến một số giá trị ngưỡng (theta) hay cần xác định một số tham số tối ưu cho bài toán qua thực nghiệm. Kết quả dự đoán có ổn định hay nhạy cảm với các giá trị này*?

Một số giá trị ngưỡng (theta) được thực nghiệm thông qua kiểm nghiệm tùy thuộc vào các bộ dữ liệu cụ thể. Ngưỡng theta lựa chọn để thực hiện thử nghiệm trong bài báo cho lại kết quả khá ổn định. Chúng tôi đã kiểm chứng khẳng định này bằng cách tìm giá trị ổn định cho các cặp pair t-test. Kết quả cho thấy giá trị ổn định luôn nhỏ hơn 0.05. Tuy vậy, điều này có còn đúng cho các bộ dữ liệu khác hay không cần phải được kiểm nghiệm để có giá trị theta thích hợp.

3. *Bài báo chỉ thử nghiệm trên 1 bộ dữ liệu thực về phim. Vậy đề xuất này có thể áp dụng cho các bộ dữ liệu thực khác không? Điều kiện trên dữ liệu về đặc trưng sản phẩm và đặc trưng người dùng thế nào thì áp dụng được*?

Đề xuất này có thể áp dụng được cho tất cả các bộ dữ liệu của lọc cộng tác hiện nay. Đối với các bộ dữ liệu có miền giá trị là các số thực trong khoảng [0, 1], ta có thể làm tương ứng với một miền nguyên xác định trong khoảng [0, 1, 2,.., g]. Khó khăn nhất trong thực nghiệm đối với mỗi bộ dữ liệu là việc xác định tập đặc trưng nội dung người dùng và sản phẩm. Thường các bộ dữ liệu chỉ cung cấp thông tin về đánh giá người dùng đối với sản phẩm mà không cung cấp thông tin các đặc trưng nội dung người dùng hoặc sản phẩm. Chính vì lý do này chúng tôi lựa chọn tập dữ liệu MovieLens (1MB) đã có đầy đủ thông tin về đánh giá người dùng và đặc trưng nội dung. Những chi tiết này chúng tôi cũng đã bổ sung trong nội dung bài báo.

II. Giải trình góp ý của phản biện 2

1. *Bài báo trình bày chưa rõ tại sao nên sử dụng thuật toán học bán giám sát để kết hợp hai phương pháp lọc trên. Thông thường thuật toán học bán giám sát được sử dụng để khắc phục vấn đề khan hiếm dữ liệu gán nhãn. Hãy làm rõ điều đó được thể hiện thế nào trong việc kết hợp hai phương pháp lọc.*

Chúng tôi hoàn toàn nhất trí với góp ý của phản biện về bản chất của thuật toán học bán giám sát. Trong quá trình nghiên cứu thực nghiệm trên các tập dữ liệu thực, chúng tôi phát hiện thấy nếu chỉ quan sát theo người dùng thì nhiều cặp người dùng gặp phải vấn đề dữ liệu thưa, nhưng khi quan sát theo sản phẩm thì tập sản phẩm chưa được người dùng đánh giá không hẳn đã thưa. Trong trường hợp này dự đoán theo phương pháp item-based có thể phát huy tác dụng tốt hơn phương pháp user-based. Ví dụ u2 trong Hình 1 gặp phải vấn đề dữ liệu thưa với tất cả tập người dùng còn lại nhưng p1, p2, p3, p4, p5 không gặp phải vấn đề dữ liệu thưa. Trong trường hợp này ta sử dụng phương pháp item-based để phát hiện nhãn dữ liệu khan hiếm cho người dùng u2. Ngược lại, khi quan sát theo sản phẩm, nhiều cặp sản phẩm phải vấn đề dữ liệu thưa, nhưng khi quan sát theo người dùng thì tập người dùng chưa đánh giá sản phẩm không hẳn đã thưa. Trong

trường hợp này dự đoán theo phương pháp user-based có thể phát huy tác dụng tốt hơn phương pháp item-based. Ví dụ p6 gặp phải vấn đề dữ liệu thưa với tất cả các sản phẩm còn lại, nhưng u1, u5, u6 không gặp phải vấn đề dữ liệu thưa. Trong trường hợp này, việc dự đoán nhãn [u6, p6] dựa vào phương pháp user-based tốt hơn phương pháp item-based.

	p_1	p_2	p_3	p_4	p_5	p_6
u_1	×	×	×	×		×
u_2	×	?	?	?	×	?
u_3		×		×		×
u_4	×		×		×	?
u_5	×	×	×	×	×	?
u_6	×	×	×	×	×	?

Hình 1. Ma trận đánh giá

Một khía cạnh quan trọng khác được phát hiện trong khi thực hiện độc lập phương pháp user-based và item-based là dự đoán theo phương pháp item-based sẽ làm cho một số cặp người dùng không gặp phải vấn đề dữ liệu thưa. Ngược lại, dự đoán theo phương pháp user-based sẽ làm cho một số cặp sản phẩm không gặp phải vấn đề dữ liệu thưa. Chính vì vậy, việc chuyển giao qua lại các nhãn phân loại hiếm giữa hai phương pháp user-based và item-based chắc chắn sẽ cải thiện được chất lượng dự đoán và hạn chế được vấn đề dữ liệu thưa của hệ tư vấn. Đây cũng là bản chất của thuật toán học bán giám sát được trình bày trong bài báo. Bán giám sát theo người dùng ta bổ sung được một số nhãn phân loại chắc chắn để chuyển giao cho quá trình bán giám sát theo sản phẩm. Ngược lại, bán giám sát theo sản phẩm ta bổ sung được một số nhãn phân loại chắc chắn chuyển giao cho quá trình bán giám sát theo người dùng. Chúng tôi đã chỉnh sửa và làm rõ nội dung này trong báo.

2. *Bài bái trình bày quá dài, quá nhiều ký hiệu và công thức rất khó theo dõi. Đề nghị tác giải viết lại bài báo một cách ngắn ngọn, tập trung làm rõ ý nghĩa của phương pháp được đề xuất cũng như các kết quả thí nghiệm chính.*

Chúng tôi xin tiếp thu và đã chỉnh sửa lại ngắn gọn và trọng tâm vào đóng góp của bài báo.

3. *Các phương pháp được so sánh với phương pháp được đề xuất đều đã rất cũ. Hầu hết được công bố trước năm 2000. Đề nghi tác giả so sánh với những phương pháp mới hơn, dược công bố gần đây hơn.*

Chúng tôi xin tiếp thu và đã bổ sung kết quả so sánh với phương pháp trong [8] của tài liệu tham khảo. Đây cũng là phương pháp có cùng hướng nghiên cứu với bài báo.

4. *Trong phần trình bày kết quả thí nghiệm độ đo MAE là gì chưa được chỉ rõ. Trong bảng 8, những giá trị nào được in đậm chưa được chú thích.*

Độ đo sai số MAE là trung bình giá trị tuyệt đối lỗi. Giá trị MAE phương pháp nào nhỏ hơn sẽ cho ta kết quả dự đoán chính xác hơn. Chúng tôi xin tiếp thu và đã bổ sung ý nghĩa của giá trị MAE vào nội dung bài báo. Những giá trị in đậm trong Bảng 8 nhằm nhấn mạnh giá trị MAE của phương pháp thấp hơn các phương pháp còn lại. Chúng tôi cũng đã chú thích lại trong nội dung bài báo.

5. *Đề nghị submit bản pdf thay vì bản word để trách việc các ký hiệu toán học bị thay đổi khi đọc trên các version khác nhau.*"

Chúng tôi nhất trí và submit bản pdf sau khi sửa đổi theo góp ý của các phản biện.

Một lần nữa chúng tôi xin trân trọng cảm ơn góp ý của các phản biện và ban biên tập.

Trân trọng,

Nhóm tác giả.