TƯ VẤN BẰNG XẾP HẠNG HÀM Ý THỐNG KÊ TRÊN DỮ LIỆU KHÔNG PHẢI NHỊ PHÂN

STATISTICAL IMPLICATIVE RATING BASED RECOMMENDATION USING NON-BINARY DATA

Phan Phương Lan[1], Nguyễn Thị Thùy Linh[1], Huỳnh Hữu Hưng[2], Huỳnh Xuân Hiệp[1]

[1]Trường Đại học Cần Thơ; {pplan, nttlinh, hxhiep}@ctu.edu.vn
[2]Trường Đại học Bách khoa - Đại học Đà Nẵng; hhhung@dut.edu.vn

Tóm tắt - Bài báo đề xuất một mô hình tư vấn lọc cộng tác dựa trên người dùng sử dụng độ đo xếp hạng hàm ý thống kê cho dữ liệu không phải nhị phân để dự đoán các xếp hạng, từ đó gợi ý cho người cần tư vấn những mục dữ liệu phù hợp. Hiệu quả của mô hình đề xuất được đánh giá qua sai số của các dự đoán (sai số tuyệt đối trung bình và căn bậc hai của sai số bình phương trung bình) và được so sánh với hiệu quả của các mô hình tư vấn lọc cộng tác dựa trên người dùng sử dụng một trong hai độ đo phổ biến Pearson và Cosine của gói recommenderlab. Kết quả thực nghiệm trên các tập dữ liệu mẫu của MovieLens và Dating cho thấy, mô hình đề xuất có sai số dự đoán thấp hơn so với các mô hình được so sánh khi số xếp hạng biết trước của người cần tư vấn nhiều hơn 2.

Từ khóa - Hệ tư vấn; độ đo xếp hạng hàm ý thống kê; lọc cộng tác dựa trên người dùng; sai số tuyệt đối trung bình; căn bậc hai của sai số bình phương trung bình.

Abstract - The paper proposes a recommendation model that uses the user based collaborative filtering approach and the statistical implicative rating measure on the non-binary data to predict the user ratings, then recommend the suitable items to users. The performance of the proposed model is evaluated by the metrics mean absolute error and root mean square error; and compared to some existing models of recommenderlab package - the user based collaborative filtering model using Cosine or Pearson. The experimental results on two datasets MovieLens and Dating show that the predictive errors of the proposed model is lower than that of compared models when the number of known ratings of user (needing the recommendation) is greater than 2.

Key words - recommender system; statistical implicative rating measure; user based collaborative filtering; mean absolute error; root mean square error.

1. Đặt vấn đề

Hệ tư vấn lọc cộng tác dựa trên người dùng (user based collaborative filtering recommender system) [1] thường sử dụng một độ đo nào đó (như Cosine, Pearson,…) để tìm những người dùng tương tự nhất với người cần tư vấn. Sau đó sử dụng thông tin xếp hạng của những người đó để dự đoán xếp hạng của người cần tư vấn cho các sản phẩm và gợi ý một danh sách sản phẩm có thể phù hợp với người này. Trong khi đó, phương pháp phân tích dữ liệu hàm ý thống kê (statistical implicative analysis) [2] thường dựa trên các độ đo như chỉ số hàm ý, cường độ hàm ý, cường độ hàm ý có entropy, hay giá trị gắn kết để phát hiện những mối quan hệ mạnh giữa các đối tượng. Vì vậy, những độ đo hàm ý thống kê có thể được sử dụng để phát triển các hệ tư vấn.

Một số nghiên cứu xây dựng hệ tư vấn dựa trên người dùng sử dụng độ đo phân tích hàm ý thống kê được trình bày trong [3], [4]. [3] đề xuất mô hình tư vấn sử dụng độ đo mới dựa trên cường độ hàm ý cho dữ liệu nhị phân; và thực hiện đánh giá mô hình này theo nhóm đo độ chính xác của các dự đoán sử dụng (như độ bao phủ - recall, độ chính xác – precision, độ đo điều hòa F1). Tuy nhiên, [3] chỉ mới xây dựng và đánh giá mô hình tư vấn cho loại dữ liệu nhị phân. [4] đề xuất mô hình tư vấn sử dụng độ đo mới dựa trên chỉ số hàm ý và thực hiện đánh giá mô hình này theo nhóm đo độ chính xác của các dự đoán sử dụng và nhóm đo độ chính xác của các dự đoán xếp hạng (như sai số tuyệt đối trung bình - MAE, căn bậc hai của sai số bình phương trung bình - RMSE). Tuy nhiên, việc đánh giá theo MAE và RMSE trong [4] là chưa thực sự phù hợp vì công thức chỉ số hàm ý mà các tác giả sử dụng là dành cho dữ liệu nhị phân, nên các giá trị xếp hạng không ở dạng nhị phân (ví dụ: các giá trị 1 hay 5) đều được tính như nhau (cùng có giá trị 1), từ đó ảnh hưởng đến việc dự đoán xếp hạng. Ngoài ra, việc đánh giá mô hình tư vấn đề xuất trong [4] chưa thực sự đầy đủ: chỉ dựa trên 25 láng giềng gần nhất và không nêu rõ số xếp hạng biết trước của mỗi người cần tư vấn trong tập dữ liệu kiểm thử.Vì vậy, để góp phần giải quyết những tồn tại vừa nêu, trong bài báo này, nhóm tác giả thực hiện xây dựng và đánh giá mô hình tư vấn cho loại dữ liệu không phải nhị phân. Cụ thể, đề xuất một mô hình tư vấn mới bằng tiếp cận lọc cộng tác dựa trên người dùng và một độ đo mới để dự đoán xếp hạng của người cần tư vấn cho một mục dữ liệu cụ thể. Độ đo mới được phát triển dựa trên cường độ hàm ý của dữ liệu không phải nhị phân.

2. Tư vấn bằng xếp hạng hàm ý thống kê trên dữ liệu không phải nhị phân

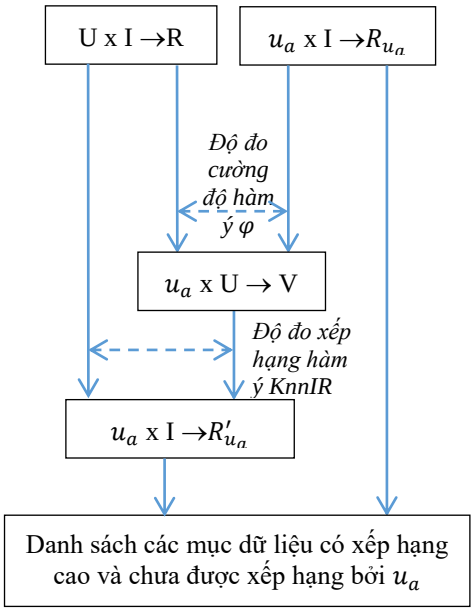
2.1. Mô hình tư vấn bằng xếp hạng hàm ý thống kê trên dữ liệu không phải nhị phân

Mô hình tư vấn bằng xếp hạng hàm ý thống kê trên dữ liệu không phải nhị phân - được phác họa như Hình 1 - sử dụng phương pháp lọc cộng tác dựa trên người dùng và độ đo xếp hạng hàm ý thống kê. Mô hình tư vấn đề xuất có:

- Một tập hữu hạn U gồm n người dùng $U = \{u_1, u_2, \ldots, u_n\}$.
- Một tập hữu hạn I gồm m mục dữ liệu (mục, item) $I = \{i_1, i_2, \ldots, i_m\}$. Mục có thể là mặt hàng, bộ phim, bài hát, v.v.
- Một ma trận xếp hạng (ma trận đánh giá – rating matrix) R lưu thông tin phản hồi của người dùng về các mục dữ liệu $R = (r_{jk})_{nxm}$ với $j = 1,..,n$ và $k = 1,..,m.r_{jk}$ có giá trị trong khoảng [0, 1]. Nếu các giá trị của R chưa ở dạng này, chúng phải được chuyển

đổi về [0, 1]. Ví dụ, nếu người dùng xếp hạng các sản phẩm theo thang từ 1 đến 5, và giá trị xếp hạng của người dùng u_j cho sản phẩm i_k là 4 thì giá trị được quy đổi là $r_{ji} = 4/5 = 0.8$.

- Một vector R_{u_a} có kích thước $m.R_{u_a}[i_k]$ với $k = 1,..,m$ lưu xếp hạng của người dùng cần tư vấn u_a cho mục dữ liệu i_k, hoặc là rỗng (NA) nếu u_a chưa xếp hạng i_k.



Hình 1. *Mô hình tư vấn bằng xếp hạng hàm ý thống kê với dữ liệu không phải nhị phân*

Mô hình tư vấn đề xuất hoạt động theo Giải thuật IIRUBCFRS với những công việc chính được thực hiện:

- Xây dựng vector V lưu cường độ hàm ý $\varphi(u_a, u_j)$ của người cần được tư vấn u_a và từng người dùng $u_j \in U$ với $j = 1..n$. $\varphi(u_a, u_j)$ được tính theo công thức (2) của Mục 2.2.
- Dự đoán xếp hạng của u_a cho các mục dữ liệu thuộc I dựa trên độ đo xếp hạng hàm ý KnnIR (K nearest neighbor Implicative Ratings) như công thức (3) của Mục 2.3.
- Lọc lại danh sách các mục dữ liệu bằng cách loại bỏ những mục đã được xếp hạng thực sự bởi u_a.
- Gợi ý cho u_a các mục dữ liệu trong danh sách đã được lọc và có giá trị xếp hạng dự đoán.

IIRUBCFRS (vector R_{u_a}; ratingmatrix R; int knn){
V = Tinhcuongdohamy(R_{u_a}, R);
R'_{u_a} = Dudoanxephang(V, R, knn);
Filteredlist = Loaibomucdaxephang(R_{u_a}, R'_{u_a});
Reclist = GoiyNmuc(Filteredlist);
return Reclist;
}

Mục tiêu của mô hình tư vấn đề xuất là dự đoán các xếp hạng sao cho sai số tuyệt đối trung bình và căn bậc hai của sai số bình phương trung bình của nó sẽ nhỏ hơn khi so với của một số mô hình lọc cộng tác dựa trên người dùng hiện có.

2.2. Độ đo cường độ hàm ý

Bên cạnh việc đề xuất hai độ đo quan trọng chỉ số hàm ý và cường độ hàm ý cho loại biến nhị phân, phương pháp phân tích hàm ý thống kê còn cập nhật công thức của hai độ đo này để áp dụng cho loại biến modal. Biến a được gọi là biến modal [2] nếu giá trị $a(i)$ của a được xác định bởi đối tượng i nằm trong khoảng [0, 1].

Chỉ số hàm ý của hai biến modal a, b được tính theo công thức (1) [2]. Trong đó, $a(i)$ và $\bar{b}(i)$ là giá trị của các biến a và $\bar{b}$ được xác định bởi đối tượng i; m_a và $m_{\bar{b}}$ là giá trị trung bình của các biến a và $\bar{b}$ tương ứng với $m_a = \frac{1}{n}\sum_{i\in E} a(i)$, $m_{\bar{b}} = \frac{1}{n}\sum_{i\in E} \bar{b}(i) = \frac{1}{n}\sum_{i\in E}(1 - b(i))$ và E gồm n đối tượng được mô tả bởi một tập hữu hạn các biến modal; và v_a và v_b là phương sai của hai biến a và b tương ứng với $v_a = \frac{1}{n}\sum_{i\in E}(a(i) - m_a)^2$,

$$v_b = \frac{1}{n}\sum_{i\in E}(b(i) - m_b)^2.$$

$$q_p(a, \bar{b}) = \frac{\frac{1}{n}\sum_{i\in E} a(i)\bar{b}(i) - m_a m_{\bar{b}}}{\sqrt{\frac{(v_a + m_a^2)(v_b + m_b^2)}{n}}} \quad (1)$$

Cường độ hàm ý của mối quan hệ a,b được xác định bởi (2) [2]. Trong đó, $q_p(a, \bar{b})$ là chỉ số hàm ý được trình bày ở trên.

$$\varphi(a, b) = \frac{1}{\sqrt{2\pi}} \int_{q_p(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}}\, dt \quad (2)$$

Để sử dụng độ đo cường độ hàm ý trong mô hình tư vấn đề xuất, ta xem các biến a và b chính là người dùng cần tư vấn u_a và một người dùng $u_j \in U$với $j = 1,..,n$ tương ứng, i chính là một mục dữ liệu $i_i \in I$. Như vậy, $\varphi(a, b)$ chính là $\varphi(u_a, u_j)$; $a(i)$ chính là $R_{u_a}[i_i]$ và $b(i)$ chính là $R[u_j, i_i]$ (tức r_{ji}). Giá cường độ hàm ý được dùng để tìm những láng giềng gần nhất của u_a. u_k là một trong knn láng giềng gần nhất của u_a nếu $\varphi(u_a, u_k)$ là một trong knn giá trị cao nhất.

Giải thuật Tinhcuongdohamy tính cường độ hàm ý giữa người dùng u_avới từng người dùng $u_j (j = 1,..,n)$ gồm các bước chính sau:

- Tính giá trị trung bình $m_a$và phương sai v_adựa trên các xếp hạng của người cần tư vấn u_a.
- Với người dùng $u_j \in U$, tính giá trị trung bình m_b và phương sai v_bdựa trên các xếp hạng của người cần tư vấn u_j, tính chỉ số hàm ý theo công thức (1) và tính cường độ hàm ý theo công thức (2) ;
- Lặp lại bước trên cho từng người dùng $u_j \in U$.

Tinhcuongdohamy(R_{u_a}, R){
mean_a = mean(R_{u_a}); var_a = var(R_{u_a});
for từng người dùng $u_j \in U$ {
mean_b = mean(R[j,]); var_b = var(R[j,]);
sum_ab_ = sum(R_{u_a}*(1-R[j,]))/n;
q_ab_ = (sum_ab_ - mean_a*(1-mean_b))/
sqrt((var_a+ mean_a^2)*(var_b+(1-mean_b)^2)/n);

```
    V[j] = 1 -pnorm(q_ab_);
    }
    return V;
}
```

2.3. Độ đo xếp hạng hàm ý thống kê

Giá trị xếp hạng (độ yêu thích) một sản phẩm i của một người dùng cần tư vấn u_a có thể gần giống như giá trị xếp hạng sản phẩm i của những người có cùng sở thích (gọi chung là các láng giềng gần nhất u_j). Việc tìm những láng giềng này có thể dựa trên cường độ hàm ý giữa hai người dùng. Tuy nhiên, nếu người dùng u_a và láng giềng u_j có mối quan hệ hàm ý mạnh nhưng độ yêu thích của u_j đối với sản phẩm i là thấp thì u_j sẽ ít ảnh hưởng đến việc xếp hạng i của u_a và ngược lại. Vì vậy, nhóm tác giả đề xuất độ đo mới KnnIR để dự đoán xếp hạng của người dùng u_a cho một mục dữ liệu i_i.

KnnIR là giá trị xếp hạng được quy đổi về khoảng [0, 1] và được xác định theo công thức (3).

$$KnnIR(u_a, i_i) = \frac{KIR(u_a, i_i)}{max_{i_l \in I} KIR(u_a, i_l)} \quad (3)$$

Trong (3), KIR (Kernel Implicative Rating) là giá trị xếp hạng gốc của người dùng u_a cho một mục dữ liệu i_i. $KIR(u_a, i_i)$ được tính bởi (4).

$$KIR(u_a, i_i) = \sum_{k=1}^{knn} \varphi(u_a, u_k) * R(u_k, i_i) \quad (4)$$

Trong đó, knn là số láng giềng gần nhất của u_a;$\varphi(u_a, u_k)$ là giá trị cường độ hàm ý của u_a với một trong knn láng giềng gần nhất $u_k \in U$ và $k = 1,..,$ knn; $R(u_k, i_i)$ là giá trị xếp hạng của một trong knn láng giềng gần nhất u_k cho mục dữ liệu i_i. $R(u_k, i_i)$ có thể được xem như trọng số của cường độ hàm ý $\varphi(u_a, u_k)$ đối với mục dữ liệu i_i.

3. Thực nghiệm

3.1. Dữ liệu và Công cụ thực nghiệm

3.1.1. Dữ liệu thực nghiệm

Bảng 1. *Thông tin chung của các tập mẫu được trích xuất từ MovieLens và Dating_4000*

Tập dữ liệu	Số người dùng	Số phim/ hồ sơ	Số xếp hạng
MovieLens	943	1,144	97,370
Dating_4000	4,000	12,476	337,830

MovieLens và Dating là hai tập dữ liệu được sử dụng trong thực nghiệm. Tập dữ liệu MovieLens [5] - lưu xếp hạng của người dùng cho các bộ phim - được thu thập thông qua trang web movielens.umn.edu trong khoảng thời gian bảy tháng. Tập dữ liệu Dating [6] lưu xếp hạng hẹn hò của người dùng cho các hồ sơ ứng viên. Tập dữ liệu MovieLens chứa: 943 người dùng; 1,664 phim và 99,392 xếp hạng với các giá trị từ 1 đến 5 và 5 là tốt nhất. Tập dữ liệu Dating có kích thước lớn nên chúng tôi lấy thông tin xếp hạng của 4,000 người dùng đầu tiên và tiến hành xây dựng ma trận xếp hạng. Kết quả, tập dữ liệu Dating_4000 có: 4,000 người dùng; 76,685 hồ sơ ứng viên và 520,732 xếp hạng với các giá trị từ 1 đến 10 và 10 là tốt nhất.

Để có thể thực nghiệm bằng mô hình tư vấn đề xuất, trước tiên, chúng tôi thực hiện tiền xử lý những tập dữ liệu này bằng cách quy đổi các giá trị xếp hạng về đoạn [0, 1]. Cụ thể, giá trị xếp hạng quy đổi được tính từ giá trị xếp hạng gốc chia cho 5 đối với tập MovieLens, chia cho 10 đối với tập Dating_4000. Bên cạnh đó, nhằm tăng tính chính xác trong việc đưa ra gợi ý, thực hiện loại bỏ những phim/hồ sơ ứng viên chỉ được xếp hạng bởi một số ít người dùng. Cụ thể, chỉ giữ lại những phim/hồ sơ được xếp hạng từ 10 người trở lên để trích xuất ra tập dữ liệu mẫu dùng trong thực nghiệm. Thông tin chung của các tập mẫu này được trình bày trong Bảng 1.

Ngoài ra, qua việc xác định phân vị người dùng (p% = a cho biết có p% số người dùng mà mỗi người đánh giá từ a mục dữ liệu trở xuống), ta có 0% số người dùng xếp hạng từ 17 phim trở xuống trên tập mẫu của MovieLens và 0% số người dùng xếp hạng từ 4 hồ sơ trở xuống trên tập mẫu của Dating_4000. Do đó, số xếp hạng biết trước (*given*) tối đa của mỗi người cần tư vấn dùng cho đánh giá hiệu quả của mô hình đề xuất là 17 trên tập mẫu của MovieLens và 4 trên tập mẫu của Dating_4000. Chi tiết về phương pháp đánh giá hệ tư vấn được trình bày trong Mục 3.2.

3.1.2. Công cụ thực nghiệm

Mô hình tư vấn đề xuất được cài đặt bằng ngôn ngữ R và tích hợp vào công cụ mà nhóm tác giả đã phát triển trong [7]. Mô hình này được đặt tên là IIMUBCF (User Based Collaborative Filtering using Modal Implicative Intensity).

Bên cạnh đó, nhóm tác giả còn sử dụng mô hình tư vấn lọc cộng tác dựa trên người dùng UBCF của gói recommenderlab [8] để so sánh với mô hình đề xuất. Mô hình UBCF dùng độ đo Cosine và Pearson – những độ đo phổ biến – để tìm những láng giềng gần nhất và xây dựng danh sách gợi ý.

3.2. Phương pháp đánh giá hệ tư vấn

Phương pháp đánh giá chéo k-tập con [9], và các độ đo sai số tuyệt đối trung bình (Mean Absolute Error - MAE), căn bậc hai của sai số bình phương trung bình (Root Mean Square Error – RMSE) [10] được sử dụng để đánh giá hiệu quả của hệ tư vấn.

Phương pháp đánh giá chéo k-tập con [9] sẽ phân tách tập dữ liệu thành k tập có kích thước bằng nhau và thực hiện k lần đánh giá sau đó lấy kết quả trung bình. Ở mỗi lần đánh giá, $(k-1)$ tập con được sử dụng làm tập huấn luyện và 1 tập con còn lại được sử dụng làm tập kiểm thử. Trong thực nghiệm này, k-tập con được đặt là 5. Tập dữ liệu kiểm thử lại được chia thành hai phần: tập dữ liệu truy vấn (queryset) và tập dữ liệu đích (targetset) có cùng kích thước. Mỗi người dùng trong tập dữ liệu truy vấn có $given$ xếp hạng được biết trước và được chọn ngẫu nhiên. Như vậy, mỗi người dùng trong tập dữ liệu đích gồm những xếp hạng còn lại. Tập dữ liệu huấn luyện cùng với tập dữ liệu truy vấn được sử dụng để dự đoán xếp hạng của những người dùng (trong tập truy vấn) cho các mục dữ liệu chưa xếp hạng. Tập dữ liệu đích cùng với các xếp hạng dự đoán được sử dụng để đánh giá hiệu quả của hệ tư vấn.

Các độ đo MAE và RMSE [10] giúp ta tính sai số của các dự đoán. Hai độ đo này có công thức tính như (5) và (6). Trong đó, $|targetset|$ là tổng số xếp hạng của tập dữ liệu đích, r_{ij} là xếp hạng thực của người dùng u_i cho mục dữ liệu i_j trong tập dữ liệu đích và r'_{ij} là xếp hạng được dự đoán của người dùng u_i cho mục dữ liệu i_j.

$$MAE = \frac{1}{|targetset|} \sum_{(i,j)\in targetset} |r_{ij} - r'_{ij}| \quad (5)$$

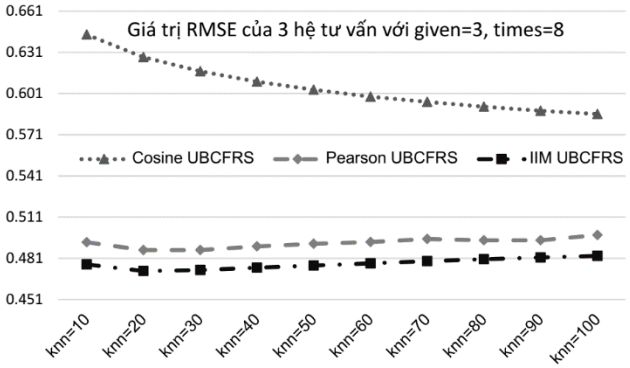
$$RMSE = \sqrt{\frac{\sum_{(i,j)\in targetset}(r_{ij} - r'_{ij})^2}{|targetset|}} \quad (6)$$

Nhóm tác giả xây dựng ba hệ tư vấn IIM UBCFRS, Cosine UBCFRS và Pearson UBCFRS, sau đó đánh giá hiệu quả của chúng. Hệ tư vấn thứ nhất sử dụng mô hình đề xuất IIMUBCF, hai hệ tư vấn còn lại sử dụng mô hình UBCF với các độ đo Cosine và Pearson tương ứng. Các thông số được sử dụng trong ba hệ tư vấn là: $given$ – số xếp hạng biết trước của mỗi người dùng trong tập truy vấn, knn – số láng giềng gần nhất, và $times$ – số lần đánh giá theo phương pháp 5-tập con. Giá trị MAE và RMSE của từng hệ tư vấn là giá trị MAE và RMSE trung bình của $times$ lần đánh giá cho từng $given$ trên cùng số láng giềng gần nhất knn. Sau đó, giá trị MAE và RMSE của ba hệ tư vấn được so sánh với nhau để đánh giá hiệu quả của chúng. Hệ tư vấn có hiệu quả tốt là hệ có sai số MAE và RMSE nhỏ.

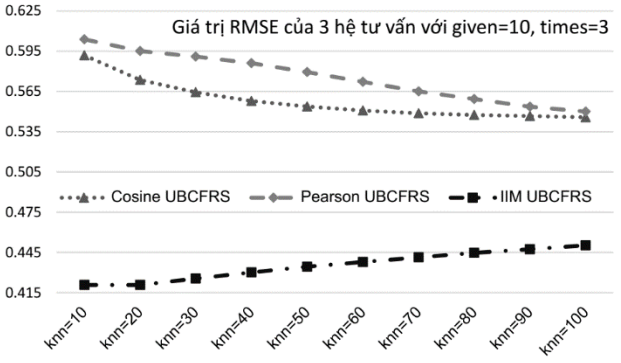
3.3. Kết quả thực nghiệm

3.3.1. Kết quả trên tập MovieLens

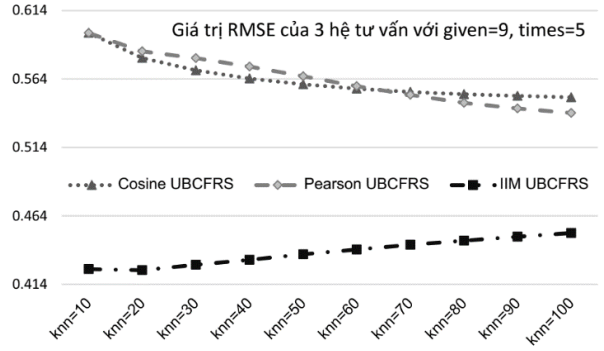
Hình 2, Hình 3 và Hình 4 hiển thị giá trị sai số RMSE của 3 hệ tư vấn có cùng số given và có cùng số láng giềng gần nhất knn thay đổi từ 10 đến 100. Cụ thể, giá trị của các tham số (given, times, knn) được sử dụng ở Hình 2 và Hình 3 là (3, 8, 10 –100), (10, 3, 10 – 100), (9, 5, 10 – 100) tương ứng. Kết quả Hình 2 và Hình 3 cho thấy, hệ tư vấn sử dụng mô hình đề xuất IIM UBCFRS có giá trị sai số thấp nhất. Kết quả Hình 4 cho thấy, giá trị RMSE của IIM UBCFRS là thấp nhất trên mọi knn được xét, của Pearson UBCFRS là cao nhất trên phân đoạn knn từ 10 đến 60, của Cosine UBCFRS là cao nhất trên phân đoạn knn còn lại. Khi thay đổi giá trị của $times$ và trên cùng số láng giềng gần nhất (knn: 10 – 100), ta nhận được kết quả tương tự như Hình 2 với từng $given$ có giá trị từ 3 đến 8, Hình 3 với từng $given$ có giá trị từ 10 đến 17 và Hình 4 với given = 9.



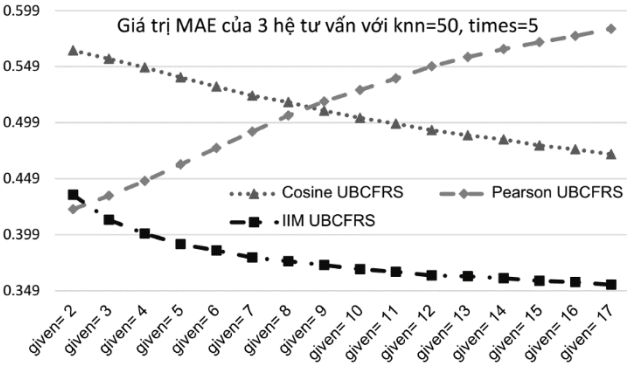
Hình 2. *Giá trị căn bậc hai của sai số bình phương trung bình của ba hệ tư vấn khi given = 3, time = 8 và knn:10 - 100*



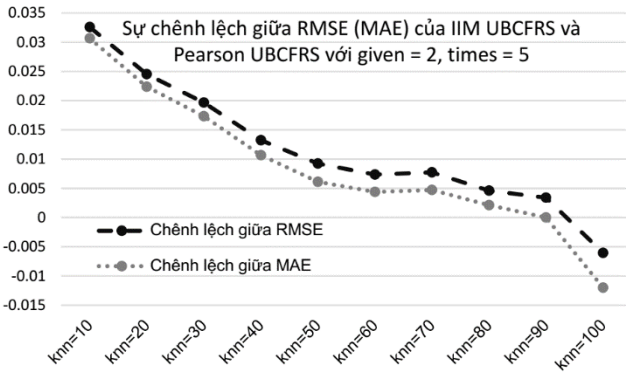
Hình 3. *Giá trị căn bậc hai của sai số bình phương trung bình của ba hệ tư vấn khi given = 10, time = 3 và knn:10 - 100*



Hình 4. *Giá trị căn bậc hai của sai số bình phương trung bình của ba hệ tư vấn khi given = 9, time = 5 và knn:10 - 100*



Hình 5. *Giá trị sai số tuyệt đối trung bình của ba hệ tư vấn khi given: 2-17, time = 5 và knn = 50*



Hình 6. *Sự chênh lệch giữa giá trị RMSE (hay giá trị MAE) của hai hệ tư vấn IIM UBCFRS và Pearson UBCFRS khi given = 2, time = 5 và knn: 10 - 100*

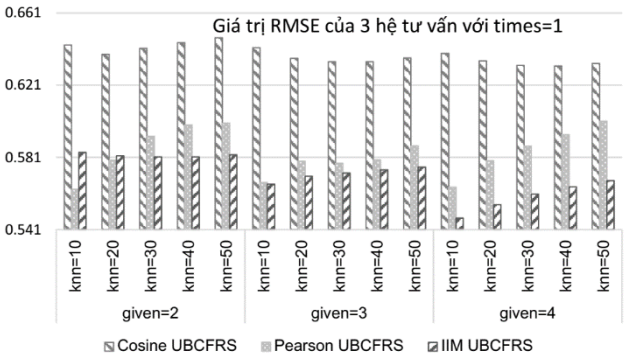
Hình 5 hiển thị giá trị MAE của 3 hệ tư vấn với số lần đánh giá times = 5 trên cùng số láng giềng gần nhất knn = 50 và có given thay đổi từ 2 đến 17. Kết quả Hình 5 cho thấy giá trị sai số MAE của hệ tư vấn sử dụng mô hình đề xuất IIM UBCFRS là thấp nhất trên hầu hết các giá trị given (từ 3 đến 17); của hệ tư vấn Cosine UBCFRS là cao nhất trên các given từ 2 đến 8; và của hệ tư vấn Pearson UBCFRS là cao nhất trên các given còn lại. Khi thay đổi giá trị của $times$ và lần lượt gán giá trị từ 10 đến 100 cho knn, ta cũng nhận được kết quả tương tự như vừa phân tích.

Trong trường hợp given = 2, giá trị RMSE và MAE của IIM UBCFRS nhỏ hơn của Cosine UBCFRS nhưng không nhỏ nhất như mong đợi. Chúng vẫn lớn hơn giá trị RMSE và MAE của Pearson UBCFRS mặc dù khoảng chênh lệch của hai hệ thống càng giảm khi knn càng tăng. Hình 6 hiển thị sự chênh lệch giá trị RMSE (giá trị MAE) của hệ thống IIM UBCFRS và hệ thống Pearson UBCFRS.

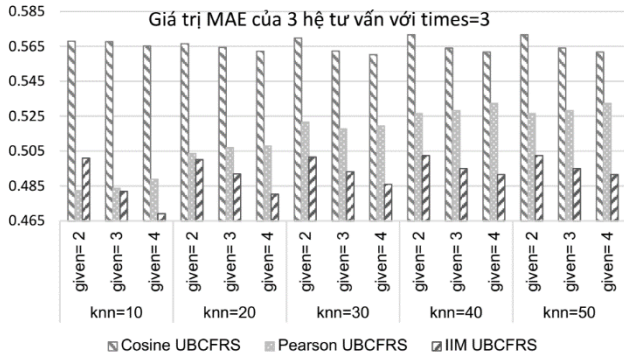
Như vậy, với mọi trường hợp ngoại trừ trường hợp given = 2, giá trị sai số dự đoán RMSE và MAE của hệ tư vấn sử dụng mô hình đề xuất IIM UBCFRS là thấp nhất. Nói cách khác, mô hình tư vấn đề xuất cho hiệu quả tốt hơn các mô hình được so sánh.

3.3.2. Kết quả trên tập Dating_4000

Hình 7 hiển thị giá trị sai số RMSE của 3 hệ tư vấn theo từng nhóm $given$ (lần lượt là 2, 3 và 4) với số knn thay đổi (từ 10 đến 50) khi times = 1.



Hình 7. *Giá trị căn bậc hai của sai số bình phương trung bình của ba hệ tư vấn khi given: 2-4, time = 1 và knn:10 - 50*



Hình 8. *Giá trị sai số tuyệt đối trung bình của ba hệ tư vấn khi given: 2-4, time = 3 và knn:10 - 50*

Kết quả Hình 7 cho thấy, hệ tư vấn sử dụng mô hình đề xuất IIM UBCFRS có giá trị sai số thấp nhất trên mọi knn khi $given$ từ 3 trở lên. Trong trường hợp given = 2, sai số dự đoán của IIM UBCFRS cao hơn của Pearson UBCFRS nhưng sẽ tiến đến nhỏ hơn khi knn tăng.

Hình 8 hiển thị giá trị sai số MAE của 3 hệ tư vấn theo từng nhóm knn (lần lượt là 10, 20, 30, 40 và 50) với $given$ thay đổi (từ 2 đến 4) khi times = 3. Kết quả Hình 8 cho thấy, giá trị sai số MAE của hệ tư vấn IIM UBCFRS là thấp nhất trên mọi $given$ khi knn từ 20 trở lên. Trong trường hợp knn = 10, sai số MAE của IIM UBCFRS cao hơn của Pearson UBCFRS khi given = 2 nhưng sẽ tiến đến nhỏ hơn khi $given$ tăng.

Do đó, với mọi knn và given >= 3, mô hình đề xuất cho hiệu quả tốt hơn mô hình được so sánh vì giá trị sai số dự đoán RMSE và MAE của hệ tư vấn sử dụng mô hình đề xuất là thấp nhất.

4. Kết luận

Bài báo đã xây dựng một mô hình tư vấn bằng tiếp cận lọc cộng tác dựa trên người dùng sử dụng độ đo xếp hạng hàm ý thống kê trên dữ liệu không phải nhị phân. Mô hình tư vấn đề xuất sử dụng độ đo mới để dự đoán xếp hạng của người cần tư vấn cho một mục dữ liệu; từ đó gợi ý cho người này những mục dữ liệu phù hợp. Giá trị của độ đo mới được dựa trên các giá trị cường độ hàm ý mạnh nhất của người cần tư vấn với những người dùng khác và thông tin xếp hạng đã biết của những người dùng đó cho mục dữ liệu đang xét. Mô hình tư vấn đề xuất được so sánh với mô hình tư vấn lọc cộng tác dựa trên người dùng UBCF của gói recommenderlab. Mô hình UBCF này sử dụng hai độ đo phổ biến Cosine/Pearson để tìm các láng giềng gần nhất và dự đoán xếp hạng. Hiệu quả của các hệ tư vấn sử dụng những mô hình (được so sánh với nhau) được đánh giá qua giá trị sai số dự đoán MAE và RMSE. Kết quả thực nghiệm trên hai tập mẫu của MovieLens/ Dating_4000 cho thấy, mô hình tư vấn đề xuất cho sai số dự đoán thấp nhất khi số xếp hạng biết trước của người cần tư vấn là từ 3 trở lên.

Tuy nhiên, khi số xếp hạng biết trước của người cần tư vấn là thấp (given = 2), hiệu quả của mô hình đề xuất chưa được như mong đợi. Cụ thể, giá trị RMSE hay MAE của mô hình đề xuất vẫn cao hơn của mô hình UBCF sử dụng độ đo Pearson mặc dù sự chênh lệch giá trị của hai mô hình càng giảm khi số láng giềng gần nhất càng tăng. Vì vậy, nhóm tác giả sẽ nghiên cứu kỹ thuật lai ghép như kết hợp mô hình dựa trên luật với mô hình đề xuất để khắc phục nhược điểm vừa nêu. Ngoài ra, việc đánh giá mô hình tư vấn có xem xét đến thứ tự của các mục dữ liệu trong danh sách gợi ý cho người cần tư vấn chưa được thực hiện trong nghiên cứu này. Do đó, trong thời gian tới, nhóm tác giả cũng sẽ thực hiện đánh giá theo hướng này để có cái nhìn đầy đủ hơn về hiệu quả của mô hình đề xuất. Bên cạnh đó, nhóm tác giả sẽ thực nghiệm so sánh mô hình đề xuất với một số mô hình tư vấn hiện có khác.

TÀI LIỆU THAM KHẢO

[1] J.B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems", *The Adaptive Web*, LNCS 4321, Springer-Verlag Berlin Heidelberg, pp. 291-324, 2007.

[2] R. Gras and P. Kuntz, "An overview of the Statistical Implicative Analysis (SIA) development", *Statistical Implicative Analysis,*

Studies in Computational Intelligence, Springer-Verlag, 127, pp. 11-40, 2008.

[3] L. P. Phan, H. H. Huynh, and H. X. Huynh, "User based Recommender Systems using Implicative Rating Measure", *International Journal of Advanced Computer Science and Applications*, Volume 10 Issue 11, 2017.

[4] Phan Quốc Nghĩa, Nguyễn Minh Kỳ, Đặng Hoài Phương, Huỳnh Xuân Hiệp, "Hệ tư vấn lọc cộng tác theo người dùng dựa trên độ đo hàm ý thống kê", *Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ Thông tin* (FAIR'9), pp. 231-239, 2016.

[5] J. Herlocker, J. Konstan, A. Borchers, J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*. Aug. 1999.

[6] L. Brozovsky and V. Petricek, "Recommender System for Online Dating Service", *Proceedings of Znalosti 2007 Conference*. 2007.

[7] L. P. Phan, N. Q. Phan, K. M. Nguyen, H. H. Huynh, H. X. Huynh, F. Guillet, "Interestingnesslab: A Framework for Developing and Using Objective Interestingness Measures", *Advances in Intelligent Systems and Computing*, Springer, 538: pp.302-311, 2017.

[8] M. Hahsler, "recommenderlab: A Framework for Developing and Testing Recommendation Algorithms", *Southern Methodist University*, 2011.

[9] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection.", *Proceedings of the14th International Joint Conference on Artificial Intelligence*, pp.1137-1143, 1995.

[10] A. Gunawardana and G. Shani, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks", *Journal of Machine Learning Research*, pp. 2935–2962, 10, 2009.

(BBT nhận bài: 22/12/2018, hoàn tất thủ tục phản biện: 24/01/2019)