**CHƯƠNG 3: PHƯƠNG PHÁP KHUYẾN NGHỊ ÁP DỤNG TRONG KHÓA LUẬN**

Sau khi tiến hành khảo sát các phương pháp khuyến nghị ở những phần trên, ở phần này chúng tôi sẽ trình bày chi tiết về phương pháp đã được hiện thực trong khóa luận này. Chương này gồm có 2 phần: phần 1 là hướng tiếp cận cắt giảm số chiều (Reduction-based), phần 2 sẽ trình bày về mô hình hồi qui được sử dụng kết hợp để dự đoán đánh giá người dùng.

**3.1. Hướng tiếp cận cắt giảm số chiều**

Ý tưởng chính của phương pháp được chúng tôi lựa chọn này là đưa bài toán đa chiều về bài toán hai chiều (User × Item) đã có lời giải. Do đó một trong những ưu điểm của phương pháp này là có thể kết hợp với bất kỳ một phương pháp khuyến nghị hai chiều nào đã trình bày ở chương 2 sau khi bước giảm chiều hoàn tất. Để hiểu phương pháp này chúng ta cùng bắt đầu quay lại với bài toán khuyến nghị hai chiều với ví dụ sau:

Với hàm dự đoán đánh giá trên, ta có D là tập chứa bộ dữ liệu <user, item, rating> cho mỗi đánh giá của từng người dùng với một đối tượng cụ thể, chúng ta có thể dự đoán bất kỳ một đánh giá nào. Tương tự ta thêm một chiều mới là thời gian thì sẽ có công thức dự đoán sẽ là:

Ở đây D là tập chứa bộ dữ liệu <user, item, time, rating> cho mỗi đánh giá cụ thể. Hàm dự đoán đánh giá trong không gian 3 chiều trên có thể biểu diễn thông qua hàm dự đoán 2 chiều như sau:

Trong đó D[Time = t](User, Item, Rating) là tập dữ liệu đánh giá trích xuất từ tập D bằng cách chọn ra những dữ liệu mà chiều Time có giá trị là t và sau đó chỉ giữ lại trường User, Item, và Rating. Nói cách khác, nếu ta biểu diễn tập dữ liệu D trong cơ sở dữ liệu quan hệ, thì D[Time = t](User, Item, Rating) là tập dữ liệu chọn lọc từ D bằng hai phép toán quan hệ: phép chọn trước và phép chiếu sau.

Cần lưu ý trong nhiều trường hợp thì tập D[Time = t](User, Item, Rating) có thể không có đủ dữ liệu đánh giá để dự đoán bằng thuật toán dự đoán hai chiều. Do đó, phương pháp tổng quát để cắt giảm số trong hệ thống khuyến nghị đa chiều có thể sẽ không sử dụng chính xác ngữ cảnh t để dự đoán đánh giá (u, i, t) mà thay vào đó là tập ngữ cảnh St (contextual segment) chứa các ngữ cảnh cao hơn hoặc bằng với t. Ví dụ, nếu muốn dự đoán điểm mà John sẽ đánh giá cho Nhà thờ Đức Bà khi đi vào ngày chủ nhật, ta có thể sẽ không lựa chọn t = “Chủ nhất” để tính toán mà thay vào đó là t = “Cuối tuần” tùy thuộc vào sự đa dạng của dữ liệu. Một cách tổng quát, ta có công thức:

Trong công thức trên ta sử dụng hàm AGGR(Rating) bởi vì một người dùng có thể đánh giá cho một đối tượng với những ngữ cảnh khác nhau, ở đây là trong những thời điểm khác nhau. Do đó, chúng ta phải kết hợp các đánh giá này lại bằng một hàm kết tập (thường là tính trung bình cộng) khi cắt giảm số chiều trong không gian khuyến nghị đa chiều.

Hướng tiếp cận cắt giảm ba chiều ở trên về hai chiều có thể được mở rộng thành một phương thức tổng quát để cắt giảm không gian khuyến nghị n-chiều thành m-chiều (với m < n). Tuy nhiên trong đề tài này chúng tôi sử dụng m = 2 để dễ dàng áp dụng các thuật toán khuyến nghị hai chiều sẵn có (chương 2).

Ta qui ước gọi hai chiều cơ bản là người dùng và đối tượng, các chiều còn lại gọi là chiều ngữ cảnh như thời gian, bạn đồng hành, thời tiết, tâm trạng, … Chúng ta cũng sử dụng thuật ngữ chuẩn trong lĩnh vực tiếp thị (marketing) để chỉ các tập dữ liệu đánh giá trích xuất theo mỗi giá trị của ngữ cảnh là “segment”, ví dụ ta có segment “Cuối tuần” trích xuất từ tập dữ liệu đánh giá D sẽ chứa tất cả các đánh giá cho những địa điểm được đi vào cuối tuần:

Tương tự segment “Cuối tuần – Bạn bè” sẽ chứa các đánh giá cho những địa điểm được đi vào cuối tuần với người yêu

Ta minh họa cách thức hoạt động của phương pháp cắt giảm số chiều bằng ví dụ sau: giả sử ta muốn dự đoán đánh giá của John cho Nhà thờ Đức Bà vào buổi sáng. Để tính , đầu tiên thuật toán sẽ loại bỏ chiều thời gian bằng cách chọn ra tập dữ liệu đánh giá mà thời gian là vào buổi sáng từ tập dữ liệu gốc D. Như vậy bài toán bây giờ được chuẩn hóa về bài toán với tập dữ liệu là các đánh giá vào buổi sáng. Sau đó chúng ta có thể áp dụng bất kỳ thuật toán dự đoán hai chiều nào như đã trình bày ở chương 2 (phương pháp chúng tôi lựa chọn là sử dụng mô hình hồi qui sẽ được trình bày chi tiết ở những phần sau). Ý tưởng của phương pháp này khá đơn giản: nếu ta muốn môt đánh giá vào buổi sáng của một người dùng cho một đối tượng cụ thể, chúng ta chỉ quan tâm tới những đánh giá đã có trước đó với thời gian vào buổi sáng.

Vấn đề tiếp theo cần được xem xét là liệu mô hình cục bộ này (mô hình khuyến nghị trên miền dữ liệu đã giới hạn theo ngữ cảnh) có tốt hơn mô hình toàn cục (mô hình khuyến nghị mà tất cả các ngữ cảnh mặc nhiên bỏ qua). Chúng ta cùng xem xét ví dụ sau:

**Ví dụ:** Ta có không gian khuyến nghị ba chiều , với X là chiều ngữ cảnh có hai giá trị là h và t. Gọi các giá trị đánh giá khi là *nt* và khi là *nh*. Giả định rằng tất cả người dùng đều có cùng sở thích như nhau (cùng thích những đối tượng với mức độ như nhau nghĩa là nt hoặc nh tùy ngữ cảnh) và .

Với giả định trên phương pháp cắt giảm số chiều luôn dự đoán các đánh giá chưa biết một cách chính xác. Có thể dễ dàng thấy điều này khi ta áp dụng phương pháp đánh giá hai chiều truyền thống, ví dụ nếu ta dùng phương pháp đánh giá sử dụng bộ nhớ:

Do đó, nếu ta sử dụng các thông tin ngữ cảnh về đối tượng đã được đánh giá, trong trường hợp segment , tất cả các đánh giá đều bằng nt, và do đó bất chấp độ đo tương quan giữa các người dùng. Tương tự với segment chúng ta luôn có . Do đó tất cả dự đoán đều chính xác với thực tế khi áp dụng phương pháp cắt giảm số chiều (mô hình cục bộ).

Ngược lại, trong mô hình toàn cục, khi áp dụng thuật toán khuyến nghị hai chiều sẽ sử dụng trộn lẫn các đánh giá *nt* và *nh* sẽ không chính xác. Tùy thuộc vào sự phân phối của các đánh giá đã có và trên một đánh giá cần dự đoán, sai số có thể chạy từ 0 (khi chỉ những dữ liệu đánh giá đúng được lựa chọn để tính toán) đến (khi chỉ những dữ liệu đánh giá sai được lựa chọ để tính toán).

Lý do mà phương pháp khuyến nghị cắt giảm số chiều chính xác hơn phương pháp khuyến nghị truyền thống là trong ví dụ trên chiều X chia miền dữ liệu ra làm hai phần riêng biệt (nhóm các đánh giá cho *X = t* là *nt* và nhóm các đánh giá cho *X = h* là *nh* với ). Tuy nhiên nếu thì chiều X sẽ không còn giá trị trong mục đích khuyến nghị (việc đánh giá trong hai trường hợp và là như nhau, nghĩa là không ảnh hưởng tới quyết định của người dùng). Trong trường hợp này, phương pháp khuyến nghị cắt giảm số chiều có độ chính xác không hơn phương pháp khuyến nghị truyền thống vì việc cắt giảm số chiều ở đây là không có ý nghĩa. Trong trường hợp tổng quát, thì khả năng X1 = X2 = … = Xn là rất thấp, do đó phương pháp khuyến nghị cắt giảm số chiều sẽ tốt hơn phương pháp khuyến nghị truyền thống trong đại đa số các trường hợp.

Bên cạnh đó ta có thể nhận thấy khả năng xảy ra trường hợp việc dự đoán trên một vài segment bằng phương pháp cắt giảm số chiều sẽ có kết quả thấp hơn phương pháp truyền thống hoặc thậm chí không thể dự đoán vì số lượng dữ liệu đánh giá trong tập segment sau khi cắt giảm là quá ít (Dữ liệu thưa thớt). Một trong những giải pháp cho vấn đề này là kết hợp phương pháp khuyến nghị cắt giảm số chiều với phương pháp khuyến nghị truyền thống sẽ được trình bày ở phần tiếp theo.

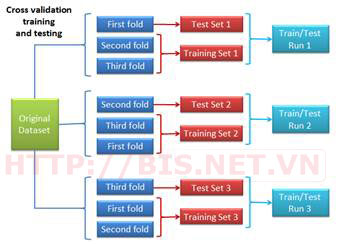
**3.2. Kết hợp phương pháp khuyến nghị cắt giảm số chiều và phương pháp khuyến nghị truyền thống.**

Trước khi đi vào phương pháp kết hợp này chúng tôi xin trình bày một vài khái niệm tiên quyết.

Để kết hợp hai phương pháp đầu tiên chúng ta cần những thước đo độ hiệu quả để so sánh phương pháp nào tốt hơn phương pháp nào trên những segment khác nhau. Một vài thước đo được sử dụng rộng rãi như: độ sai lệch trung bình tuyệt đối (MAE – mean absolute error), độ sai lệch trung bình bình phương (MSE – mean squared error), sự tương quan giữa dự đoán và thực tế (F-measure).

Trong phần này, chúng tôi sử dụng một hàm đo độ hiệu quả trừu tượng là sử dụng cho một thuật toán khuyến nghị A với tập dữ liệu huấn luyện X và đánh giá cho tập dữ liệu Y (có sẵn), với . Với mỗi , gọi *d.R* là giá trị đánh giá thực cho điểm d, *d.RA,X* là đánh giá được dự đoán bằng thuật toán A (bất kỳ thuật toán khuyến nghị nào như đã trình bày ở chương 2) trên tập dữ liệu huấn luyện X cho điểm d. Hàm được định nghĩa như một hàm thống kê trên hai tập dữ liệu đánh giá và . Ví dụ độ sai lệch trung bình tuyệt đối được định nghĩa như sau:

Như đã đề cập ở trên, trong thì , nói cách khác tập dữ liệu huấn luyện và tập dữ liệu đánh giá phải riêng biệt với nhau. Trong các nghiên cứu người ta thường sử dụng kỹ thuật “N-fold Cross Validation“, dữ liệu gốc sẽ được chia thành n phần bằng nhau, và quá trình huấn luyện/đánh giá cho model thực hiện lặp lại n lần. Tại mỗi lần huấn luyện/đánh giá, 1 phần dữ liệu dùng để đánh giá và (n-1) phần còn lại dùng để huấn luyện. Ví dụ sau đây cho n = 3, dữ liệu ban đầu chia làm 3 phần, và có 3 lần thực hiện quá trình huấn luyện/đánh giá. Lần thứ nhất, phần 1 làm tập đánh giá, 2 phần còn lại làm tập huấn luyện. Lần thứ hai, phần 2 làm tập đánh giá. Lần thứ ba, tới phần 3 làm tập đánh giá …



Hình: Cross Validation với n = 3

Sau ba pha chạy như trên, ở mỗi pha sẽ tính toán lỗi (dựa trên số phần tử trong tập Test Set có được phân lớp chính xác hay không). Từ đó cho ra chỉ số đánh giá độ chính xác trung bình cho tập dữ liệu huấn luyện ban đầu.

Áp dụng kỹ thuật Cross Validation trên tập dữ liệu đánh giá đã có *T* ta sẽ có các tập dữ liệu huấn luyện và đánh giá *Xi* và *Yi* ( i=1, 2, …), với và . Ta có thể dự đoán đánh giá cho *d.R* bằng cách tính trung bình cộng các điểm dự đoán trên các tập huấn luyện như sau:

, với

Ta có độ đo hiệu quả của thuật toán A trên tập dữ liệu T là

Bây giờ chúng ta sẽ tiếp tục với phương pháp kết hợp giữa phương pháp cắt giảm số chiều và phương pháp khuyến nghị truyền thống để giải quyết vấn đề đã nêu ở phần trước. Phương pháp này gồm có hai pha: đầu tiên chúng tôi sử dụng các đánh giá của người dùng đã biết để chọn ra những segment vượt trội so với phương pháp truyền thống (áp dụng thuật toán trên toàn bộ dữ liệu). Sau đó, để dự đoán một đánh giá, chúng tôi sẽ chọn ra segment phù hợp nhất với điều kiện ngữ cảnh và áp dụng thuật toán khuyến nghị hai chiều trên segment này. Cụ thể như sau:

* Pha thứ nhất được chạy “offline” và gồm có ba bước:
* Bước 1: Tìm tất cả các segment với số lượng đánh giá lớn hơn một ngưỡng định sẵn.
* Bước 2: Từ tập segment S trên, ta chạy thuật toán A trên S và tính toán độ hiệu quả với kỹ thuật lấy mẫu cross-validation. Ta cũng chạy thuật toán A trên toàn bộ tập dữ liệu T và tính được độ hiệu quả trên tập dữ liệu đánh giá S. Sau đó so sánh hai kết quả này để quyết định xem phương pháp nào hiệu quả hơn với cùng tập dữ liệu đánh giá S và chỉ giữ lại những segment khi áp dụng phương pháp cắt giảm số chiều đạt hiệu quả cao hơn này.
* Bước 3: Loại bỏ những segment có được sau bước 2 nếu trong tập segment này cũng tồn tại một segment Q tổng quát hơn nó và có độ hiệu quả cũng cao hơn.

|  |  |
| --- | --- |
| **Đầu vào:** | |
| ***T*** | Tập dữ liệu đánh giá trong không gian khuyến nghị đa chiều. |
| *RA,T* | Hàm dự đoán đánh giá dựa trên thuật toán A và tập dữ liệu huấn luyện T. |
| *µ* | Hàm đo độ hiệu quả |
| *N* | Số đánh giá có sẵn nhỏ nhất cần có của một segment hợp lệ |
|  | |
| **Đầu ra:** | |
|  | Tập segment mà bằng phương pháp cắt giảm số chiều dựa trên thuật toán A có hiệu quả cao hơn thuật toán A thuần túy. |
|  | |
| **Thuật toán:** | |
| 1. Tìm tập *SEGM(T)* là tập các segment có số lượng đánh giá đã có lớn hơn hoặc bằng N. 2. Với mỗi segment tính *µA,S(S)* và *µA,T(S)*, và chỉ giữ lại những segment mà *µA,S(S)* tốt hơn *µA,T(S).* 3. Với những segment còn lại trong *SEGM(T)* sau bước 2, loại bỏ tất cả những segment S nếu tồn tại một segment Q mà và *µA,Q(Q)* tốt hơn *µA,S(S).* Tập segment được giữ lại sau cùng là cần tìm. | |

* Pha thứ hai là bước chạy trực tuyến (online) để dự đoán một đánh giá của một người dùng cho một đối tượng với điều kiện ngữ cảnh cho trước. Gọi *d* là điểm cần dự đoán đánh giá, đầu tiên ta duyệt qua các segment ứng viên theo thứ tự giảm dần độ hiệu quả, sau đó chọn ra segment đầu tiên mà *d* thuộc về (theo ngữ cảnh). Nếu *d* không nằm trong bất cứ segment ứng viên nào, ta sẽ sử dụng thuật toán *A* trên dữ liệu toàn cục T để dự đoán đánh giá cho *RA,T(d)*. Ngược lại nếu tìm được segment *Sj* phù hợp thì kết quả trả về sẽ là .

|  |  |
| --- | --- |
| **Đầu vào:** | |
|  | Tập segment S1 đến Sk được sắp xếp theo thứ tự giảm dần độ hiệu quả *µ*, nghĩa là . |
| *d* | Điểm d cần dự đoán đánh giá. |
|  | |
| **Đầu ra:** | |
| *d.R* | Đánh giá dự đoán cho d |
|  | |
| **Thuật toán:** | |
| 1. *j = 0.* 2. Nếu j = 0 thì //d không thuộc bất kỳ segment Si nào.   Ngược lại thì | |

Dễ dàng thấy được bằng phương pháp kết hợp này chúng ta sẽ luôn có được kết quả chính xác hơn hoặc bằng với phương pháp truyền thống (hai chiều) bởi vì nếu phương pháp cắt giảm số chiều đạt hiệu quả tốt chúng ta sẽ sử dụng nó còn không sẽ sử dụng phương pháp truyền thống.