KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN

CHƯƠNG TRÌNH CHÍNH QUY

Kết hợp các đối tượng và đặc trưng lân cận cho bài toán phân lớp đa nhãn

Giáo viên hướng dẫn: GS.TS Lê Hoài Bắc

Nhóm thực hiện:

- Trần Xuân Quý – 18120231
- Trần Hữu Chí Bảo – 18120288

Nội dung

- Giới thiệu đề tài
- Mục tiêu đề tài
- Cơ sở lý thuyết
- Đề xuất giải pháp
- Thiết kế, thực nghiệm mô hình với bài toán phân lớp đa nhãn
- Kết quả thử nghiệm
- Kết luận và hướng phát triển

Giới thiệu đề tài

Giới thiệu đề tài

Giới thiệu đề tài

Thách thức:

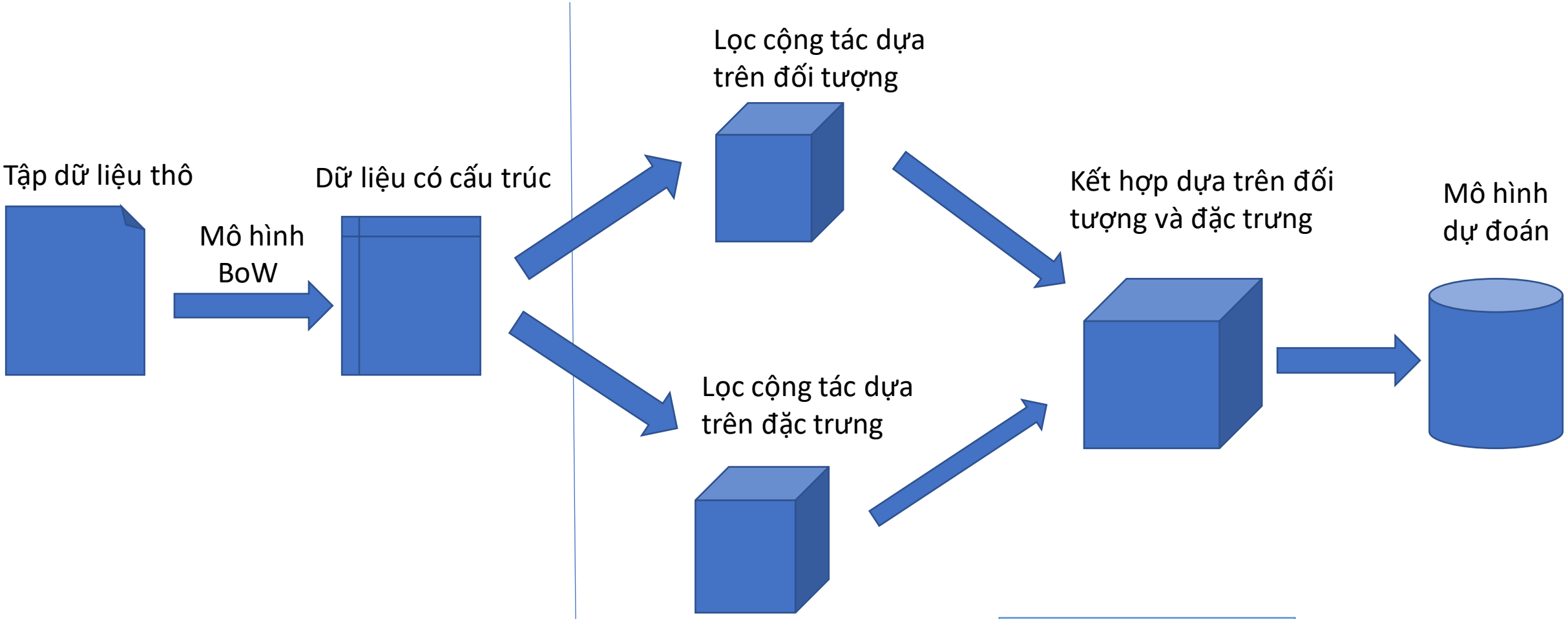
- Mối quan hệ giữa các nhãn với nhau.
- Sự chênh lệch nhãn.

Mục tiêu đề tài

- Tìm mối quan hệ sự phụ thuộc giữa các nhãn.
- Triển khai khắc phục hạn chế ở mô hình cơ bản.
- Nghiên cứu đặc điểm, cách hoạt động của phương pháp đề xuất.
- Cải thiện chất lượng của mô hình.

Nội dung đề tài

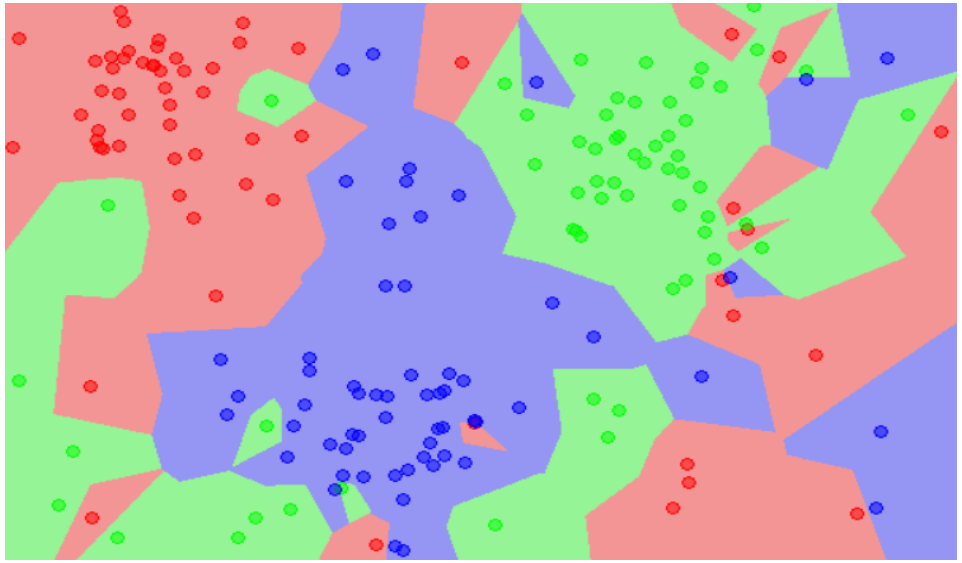
Mô hình khai thác dữ liệu



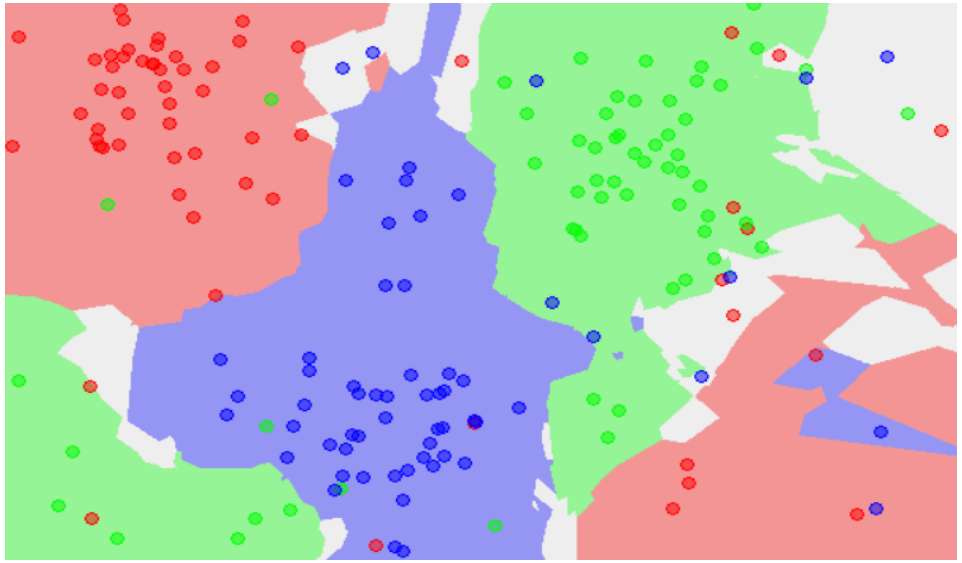
Cơ sở lý thuyết

Thuật toán KNN

- K lân cận gần nhất (KNN) là thuật toán học có giám sát (supervised learning).
- Ý tưởng của KNN là tìm ra output của dữ liệu dựa trên thông tin của những dữ liệu training gần nó nhất.



Phân loại 1NN



Phân loại 5NN

Lọc cộng tác

Lọc cộng tác dựa trên đối tượng

Ý tưởng xác định mức độ quan tâm của một user tới một item dựa trên các users khác gần giống với user này.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	4	?	?	0	?	2	?
i_2	?	4	1	?	?	1	1
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5
$\bar{u}_j$	3.25	2.75	2.5	1.33	2.5	1.5	3.33

a) Original utility matrix $\mathbf{Y}$ and mean user ratings.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	1.75	2.25	-0.5	-1.33	-1.5	0	0
i_1	0.75	0	0	-1.33	0	0.5	0
i_2	0	1.25	-1.5	0	0	-0.5	-2.33
i_3	-1.25	-0.75	0.5	2.67	1.5	0	0.67
i_4	-1.25	-2.75	1.5	0	0	0	1.67

b) Normalized utility matrix $\bar{\mathbf{Y}}$.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
u_0	1	0.83	-0.58	-0.79	-0.82	0.2	-0.38
u_1	0.83	1	-0.87	-0.40	-0.55	-0.23	-0.71
u_2	-0.58	-0.87	1	0.27	0.32	0.47	0.96
u_3	-0.79	-0.40	0.27	1	0.87	-0.29	0.18
u_4	-0.82	-0.55	0.32	0.87	1	0	0.16
u_5	0.2	-0.23	0.47	-0.29	0	1	0.56
u_6	-0.38	-0.71	0.96	0.18	0.16	0.56	1

c) User similarity matrix S.

Thực hiện dự đoán điểm số của $u1$ dành cho $i1$ với số lân cận gần nhất k = 2

Độ tương đồng cosine là một cách thể hiện sự giống nhau giữa 2 dãy số.

$$cosine\_similarity(A,B) = cos(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	1.75	2.25	-0.5	-1.33	-1.5	0.18	-0.63
i_1	0.75	0.48	-0.17	-1.33	-1.33	0.5	0.05
i_2	0.91	1.25	-1.5	-1.84	-1.78	-0.5	-2.33
i_3	-1.25	-0.75	0.5	2.67	1.5	0.59	0.67
i_4	-1.25	-2.75	1.5	1.57	1.56	1.59	1.67

d) $\hat{\mathbf{Y}}$

Predict normalized rating of u_1 on i_1 with $k = 2$

Users who rated i_1 : $\{u_0, u_3, u_5\}$

Corresponding similarities: {0.83, -0.40, -0.23}

$\Rightarrow$ most similar users: $\mathcal{N}(u_1, i_1) = \{u_0, u_5\}$

with **normalized ratings** {**0.75, 0.5**}

$\Rightarrow \hat{y}_{i_1,u_1} = \frac{0.83*0.75+(-0.23)*0.5}{0.83+|-0.23|} \approx 0.48$

e) Example

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	1.68	2.70
i_1	4	3.23	2.33	0	1.67	2	3.38
i_2	4.15	4	1	-0.5	0.71	1	1
i_3	2	2	3	4	4	2.10	4
i_4	2	0	4	2.9	4.06	3.10	5

f) Full $\mathbf{Y}$

Điểm dự đoán được xác định theo công thức:

$$\hat{y}_{i_j,u_k} = \frac{\sum_{u_l \in KNN(u_k)} r_{j,l} \times sim(u_k, u_l)}{\sum_{u_l \in KNN(u_k)} |sim(u_k, u_l)|}$$

Lọc cộng tác dựa trên đặc trưng

Ý tưởng Sử dụng trung bình của từng đặc trưng để chuẩn hóa giá trị thay cho trung bình với từng người dùng và ma trận tương đồng tính theo những đặc trưng.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6		
i_0	5	5	2	0	1	?	?	→	2.6
i_1	4	?	?	0	?	2	?	→	2
i_2	?	4	1	?	?	1	1	→	1.75
i_3	2	2	3	4	4	?	4	→	3.17
i_4	2	0	4	?	?	?	5	→	2.75

a) Original utility matrix $\mathbf{Y}$ and mean item ratings.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	2.4	2.4	-.6	-2.6	-1.6	0	0
i_1	2	0	0	-2	0	0	0
i_2	0	2.25	-0.75	0	0	-0.75	-0.75
i_3	-1.17	-1.17	-0.17	0.83	0.83	0	0.83
i_4	-0.75	-2.75	1.25	0	0	0	2.25

b) Normalized utility matrix $\bar{\mathbf{Y}}$.

	i_0	i_1	i_2	i_3	i_4
i_0	1	0.77	0.49	-0.89	-0.52
i_1	0.77	1	0	-0.64	-0.14
i_2	0.49	0	1	-0.55	-0.88
i_3	-0.89	-0.64	-0.55	1	0.68
i_4	-0.52	-0.14	-0.88	0.68	1

c) Item similarity matrix $\mathbf{S}$.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	2.4	2.4	-.6	-2.6	-1.6	-0.29	-1.52
i_1	2	2.4	-0.6	-2	-1.25	0	-2.25
i_2	2.4	2.25	-0.75	-2.6	-1.20	-0.75	-0.75
i_3	-1.17	-1.17	-0.17	0.83	0.83	0.34	0.83
i_4	-0.75	-2.75	1.25	1.03	1.16	0.65	2.25

d) Normalized utility matrix $\bar{\mathbf{Y}}$.

Thực hiện dự đoán điểm số của u_6 dành cho i_0 với số lân cận gần nhất k = 2:

$$\hat{y}_{i_j,u_k} = \frac{\sum_{u_l \in KNN(u_k)} r_{j,l} \times sim(u_k, u_l)}{\sum_{u_l \in KNN(u_k)} |sim(u_k, u_l)|}$$

Trong đó:

- $\hat{y}_{i_j,u_k}$: điểm dự đoán cho sản phẩm i_j của đối tượng u_k.
- u_l: người dùng có độ tương đồng cao nhất.
- $r_{j,l}$: đánh giá cho sản phẩm i_j của người dùng u_l
- $sim(u_k, u_l)$: độ tương đồng giữa người dùng u_k và u_l.

Mô hình kết hợp K lân cận gần nhất dựa trên đối tượng và dựa trên đặc trưng

Linear Combination of Instance-based and Feature-based k-nearest neighbors

Các thành phần

K-NN dựa trên đối tượng.

- Độ tương đồng:

$$sim_{INS} = \frac{x_q \cdot x_i}{\|x_q\|_2 \cdot \|x_i\|_2}$$

- Dự đoán:

$$\hat{y}_{q,j}^{INS} = \frac{\sum_{x_i \in kNN(x_q)} y_{i,j} \cdot sim_{INS}(x_q, x_i)^{\alpha}}{\sum_{x_i \in kNN(x_q)} sim_{INS}(x_q, x_i)^{\alpha}}$$

K-NN dựa trên đặc trưng.

- Độ tương đồng:

$$sim_{FL} = \frac{f_i \cdot l_j}{\|f_i\|_2 \cdot \|l_j\|_2}$$

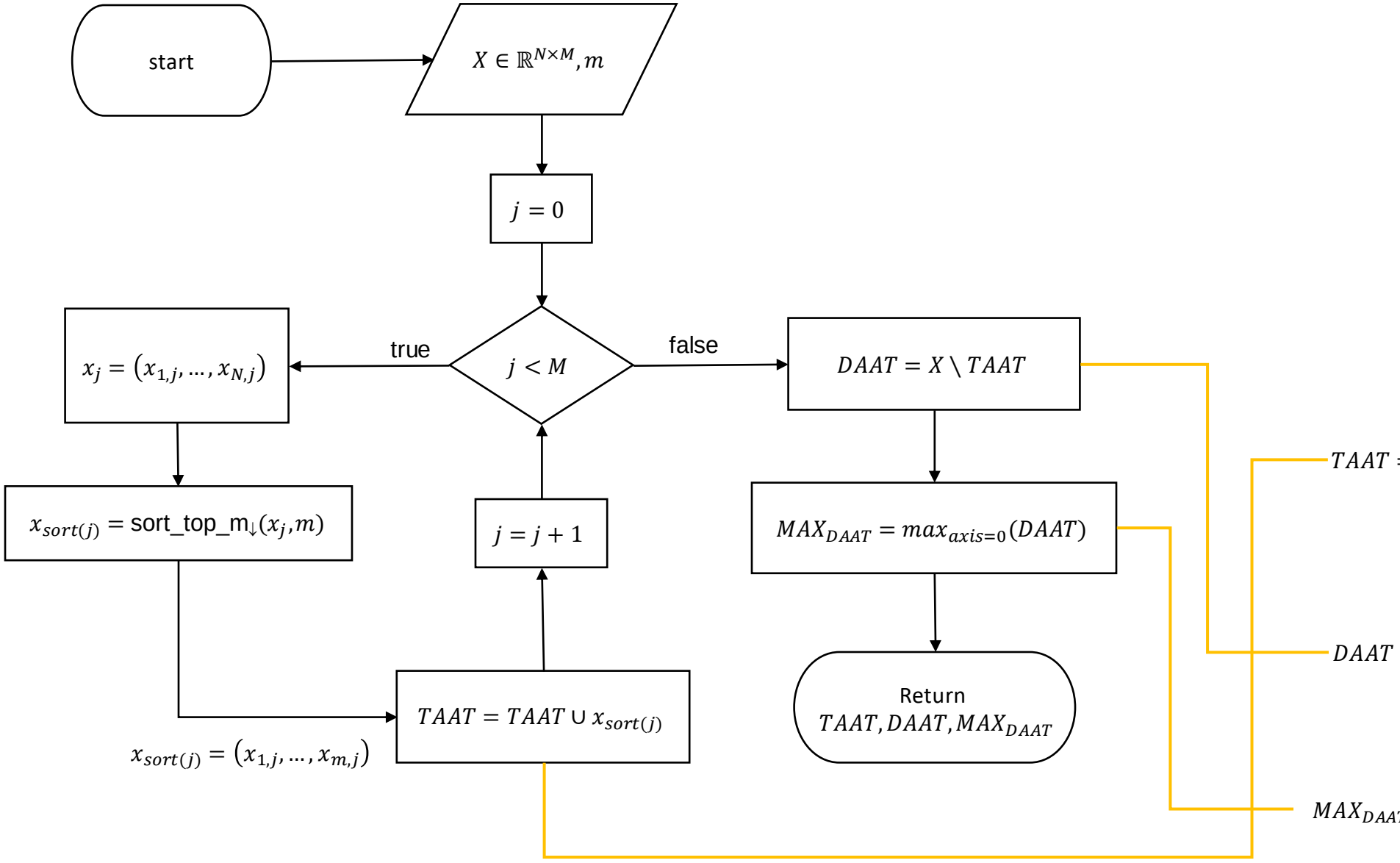
- Dự đoán:

$$\hat{y}_{q,j}^{FL} = \frac{\sum_{i=1}^{M} x_{q,i} \cdot sim_{FL}(f_i, l_j)^{\beta}}{\sum_{i=1}^{M} x_{q,i}}$$

- Sử dụng độ tương đồng thay cho độ đo khoảng cách.
- Kết quả tương đồng là trọng số dự đoán.
- Tổng hợp kết quả:

$$\hat{Y} = \lambda \hat{Y}_{INS} + (1 - \lambda)\hat{Y}_{FL}$$

Partition(X, m)



vd $m = 1$

x	f_1	f_2	f_3	f_4	f_5
0	0.71	0.57	0	0.29	0.29
1	0.75	0	0.6	0.3	0
2	0.36	0	0.18	0.54	0.73
3	0.24	0	0	0.97	0.75
4	0	0.89	0.44	0	0

$TAAT =$

x	f_1	f_2	f_3	f_4	f_5
1	0.75	0	0.6	0.3	0
3	0.24	0	0	0.97	0.75
4	0	0.89	0.44	0	0

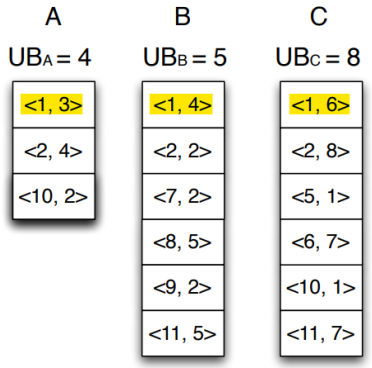
$DAAT =$

x	f_1	f_2	f_3	f_4	f_5
0	0.71	0.57	0	0.29	0.29
2	0.36	0	0.18	0.54	0.73

$MAX_{DAAT} =$

f_1	f_2	f_3	f_4	f_5
0.71	0.57	0.18	0.54	0.73

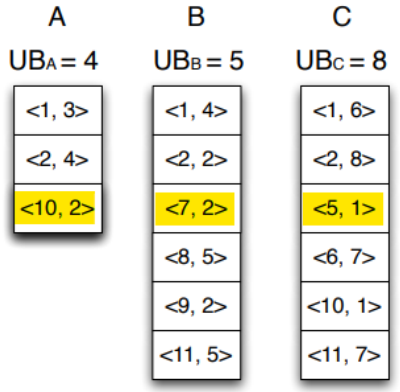
Document-at-a-time + Weak AND



Posting list và biên trên cho đặc trưng A, B và C.

Heap	
docid	$score(d, Q)$
1	13 (θ)
2	14

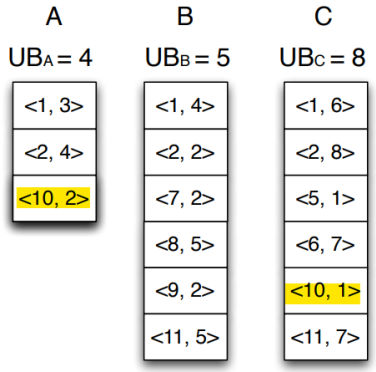
	C	B	A
p	1	2	3
docid	5	7	10



$k = 2$ đối tượng có điểm số cao nhất trong Heap sau khi xử lý đối tượng 1 và 2.

WAND duyệt mảng tổng hợp và chọn ứng viên:
p=1: $UB_C=8<\theta=13$.
p=2: $UB_C+UB_B=8+5=\theta=13$.
p=3:
$UB_C+UB_B+UB_A=8+5+4>\theta=13$.
Xác định ứng viên là đối tượng 10

Điểm tương đồng của đối tượng 10 $score(10, Q) = 3 < \theta$ nên không thêm vào Heap.



Nguồn: Fontoura, M., Josifovski, V., Liu, J., Venkatesan, S., Zhu, X., & Zien, J. (2011). Evaluation strategies for top-k queries over memory-resident inverted indexes. *Proceedings of the VLDB Endowment*, *4*(12), 1213-1224.

Document-at-a-time + Weak AND

dataset	train	test	N_{TAAT}	$prod_{DAAT}$
medical	333	645	312	556
corel5k	5000	500	4500	500
bibtex	4880	2515	4880	0
delicious	12920	3185	12631	2440

Số cặp đối tượng cần tinh tương đồng khi áp dụng DAAT + Weak AND so với không áp dụng (train * test)

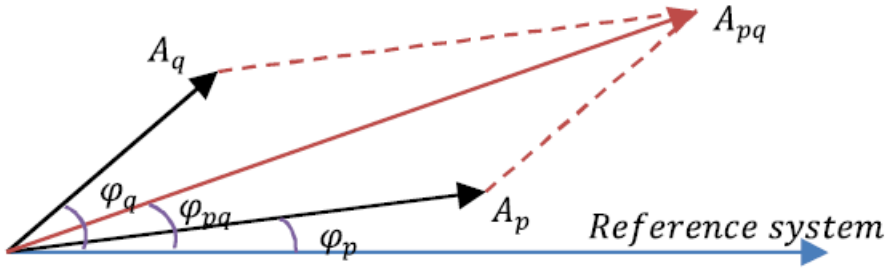
Độ tương đồng cộng hưởng

- Dựa trên hiện tượng cộng hưởng trong dao động điều hòa đơn giản.
  - $\left|\varphi_p - \varphi_q\right| \nearrow\swarrow \ A_{pq}$
- Công thức:

$$RES(u,v) = \sum_{I_{u,v}} C(u,v,k_1) * D(u,v,k_2,k_3) * J(u,v,k_4)$$

  - Giới hạn giá trị về miền $[0,1)$:

$$sim_{RES}(u,v) = \frac{\tan^{-1} RES(u,v)}{0.5\pi}$$



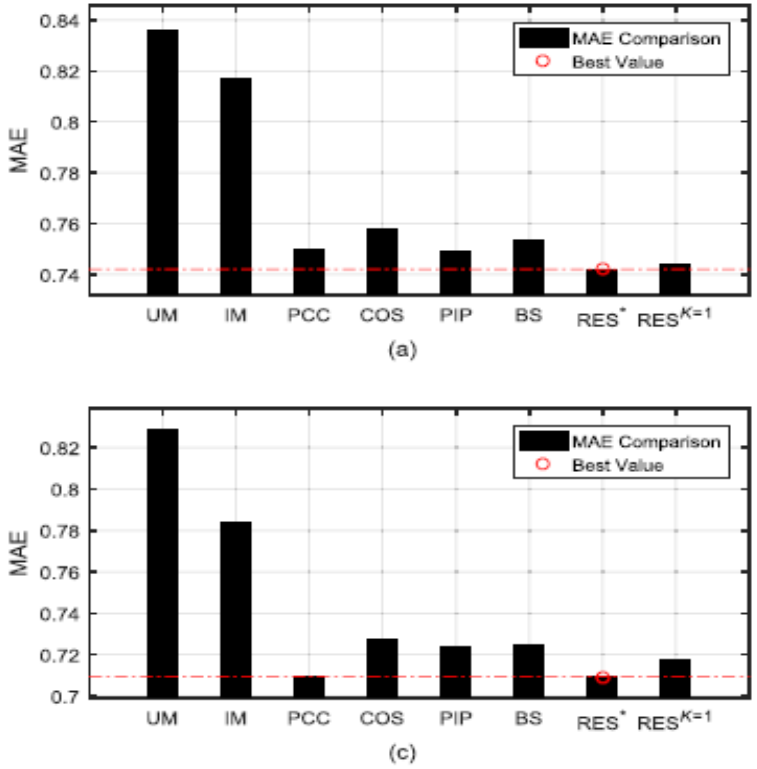
Ví dụ 2 dao động điều hòa và biên độ cộng hưởng

Nguồn ảnh: TAN, Zhenhua; HE, Liangliang. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle.

Độ tương đồng cộng hưởng

Problem	Examples			PCC	COS	PIP	RES_Eq3	RES_Eq4
	ID	Vector u	Vector v				$\mathbb{K}$ = [1,1,1,1]	$\mathbb{K}$ = [1,1,1,1]
Equal-ratio	1	[1,2,1]	[1.5,3,1.5]	1.0	1.0	6015.4719	10.3781	0.9388
	2	[1,2,1]	[2.5,5,2.5]	1.0	1.0	2029.6406	2.2512	0.7339
	3	[1,2,1,1,4]	[0.5,1,0.5,0.5,2]	1.0	1.0	14913.2172	19.8975	0.9680
Unequal-length	4	[1,2]	[1,2]	1.0	1.0	5690.250	13.8312	0.9541
	5	[1,2,4,5,5,5,5,5]	[1,2,5,5,5,5,5,5]	0.9765	0.9971	28257.870	65.4555	0.9903
Flat-value	6	[1,1,1]	[1,1,1]	NaN	1.0	11911.860	25.6892	0.9752
	7	[1,1,1]	[3,3,3]	NaN	1.0	441.0	1.1164	0.5350
	8	[1,1,1]	[5,5,5]	NaN	1.0	3.0	0.0412	0.0262
Opposite-value	9	[1,5,1]	[5,1,5]	-1.0	0.4042	3.0	0.0125	0.0080
	10	[2,4,2]	[4,2,4]	-1.0	0.8165	75.0	0.2364	0.1478
	11	[2,3,2]	[4,3,4]	-1.0	0.9469	302.0	1.8632	0.6864
Single-value	12	[1]	[1]	NaN	1.0	5285.250	12.1825	0.9479
	13	[1]	[3]	NaN	1.0	147.0	0.5684	0.3291
	14	[1]	[5]	NaN	1.0	1.0	0.0137	0.0087
Cross-value	15	[1,5]	[5,1]	-1.0	0.3846	2.0	0.0002	0.0001
	16	[1,4]	[4,2]	-1.0	0.4706	31.0	0.0757	0.0481
	17	[2,4]	[3,2]	-1.0	0.8682	324.0	0.9626	0.4879
	18	[5,1]	[5,4]	1.0	0.8882	2585.250	3.6253	0.8287
	19	[5,1]	[5,2]	1.0	0.9833	3137.250	5.2020	0.8791
	20	[5,2]	[4,1]	1.0	0.9908	1536.0	1.9971	0.7045

Kết quả tương đồng của một số véc-tơ dữ liệu ví dụ trên các độ tương đồng



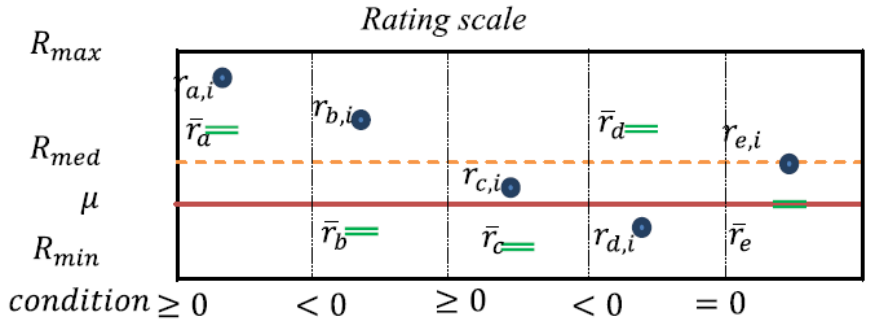
So sánh kết quả dự đoán đánh giá với các độ tương đồng ở tập Movielens-100k và Movielens-1M

Nguồn ảnh: TAN, Zhenhua; HE, Liangliang. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle

Độ tương đồng cộng hưởng

Độ nhất quán:

- Ảnh hưởng bởi $\varphi(u) - \varphi(v)$
- φ: thái độ, ý kiến chủ quan trong đánh giá:
  - $(r_{u,i} - R_{med})$: đánh giá cá nhân so với toàn tập dữ liệu.
  - $(\bar{r}_u - \mu)$: điểm trung bình người dùng so với tất cả sản phẩm.



Ví dụ các trường hợp của $condition = (r_{u,i} - R_{med}) \times (\bar{r}_u - \mu)$:
$condition \geq 0$: $\varphi \downarrow$
$condition < 0$: $\varphi \uparrow$

Nguồn ảnh: TAN, Zhenhua; HE, Liangliang. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle

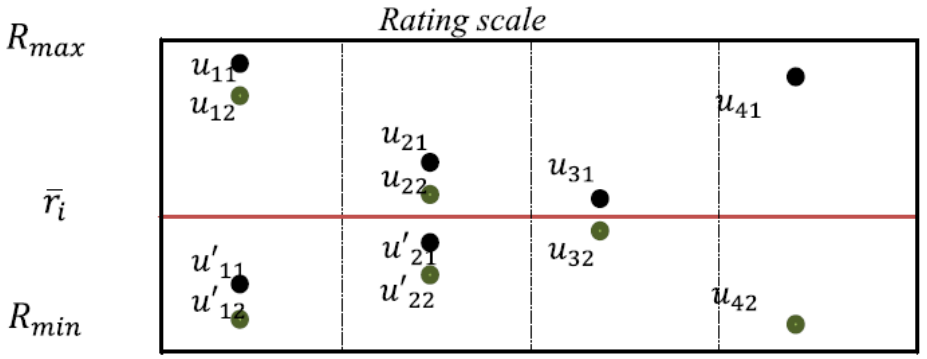
Độ tương đồng cộng hưởng

Khoảng cách:

- Thể hiện ý kiến chung về sản phẩm.
- 2 trường hợp:
  - $\left(r_{u,i} - \bar{r}_i\right)\left(r_{v,i} - \bar{r}_i\right) \geq 0$
  - $\left(r_{u,i} - \bar{r}_i\right)\left(r_{v,i} - \bar{r}_i\right) < 0$

Hệ số Jaccard:

$$J(u, v, k_4) = \frac{\left|I_{u,v}\right|}{|I_u| + |I_v| - \left|I_{u,v}\right|}$$

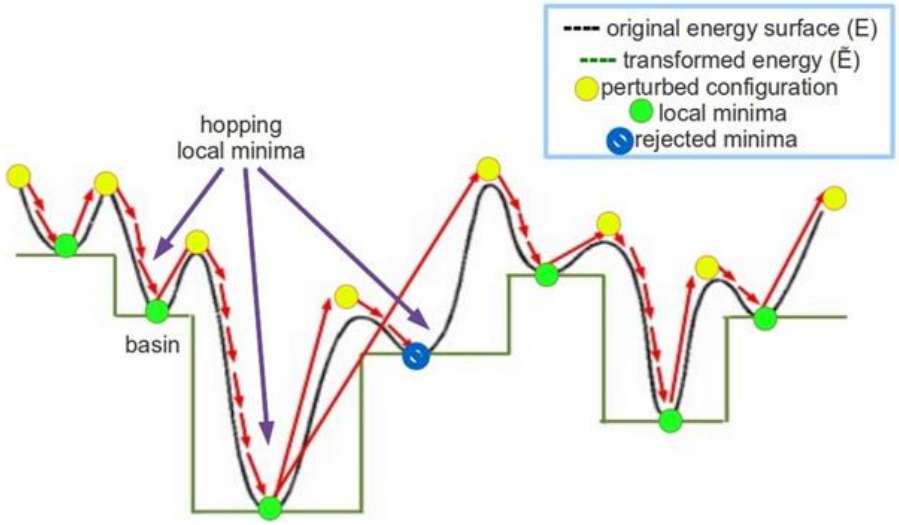


ví dụ về khác biệt trong đánh giá:
Yếu tố ảnh hưởng khi $\left(r_{u,i} - \bar{r}_i\right)\left(r_{v,i} - \bar{r}_i\right)$:
≥ 0: $\left|r_{u,i} - r_{v,i}\right|$ và $\left|r_{u,i} - \bar{r}_i\right| + \left|r_{v,i} - \bar{r}_i\right|$
< 0: $\left|r_{u,i} - r_{v,i}\right|$

Nguồn ảnh: TAN, Zhenhua; HE, Liangliang. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle.

Độ tương đồng cộng hưởng

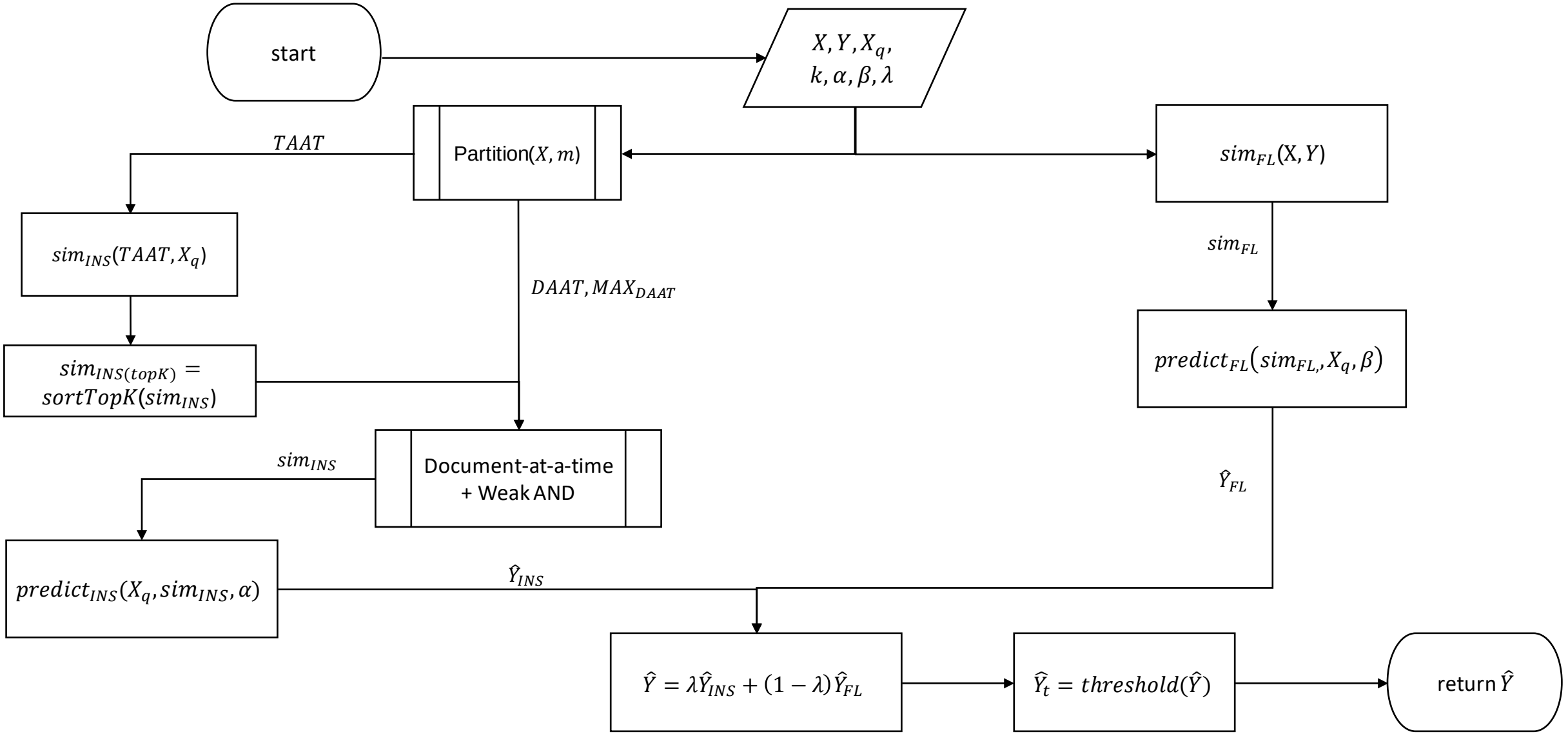
- Tìm kiếm tham số $k_1 \sim k_4$ tối ưu: thuật toán Basin-hopping.
  - Chọn ngẫu nhiên bộ tham số
  - Tối ưu cục bộ
- Hàm mục tiêu: dự đoán dựa trên đối tượng trên độ đo xác định



Tổng quan hoạt động của Basin-hopping

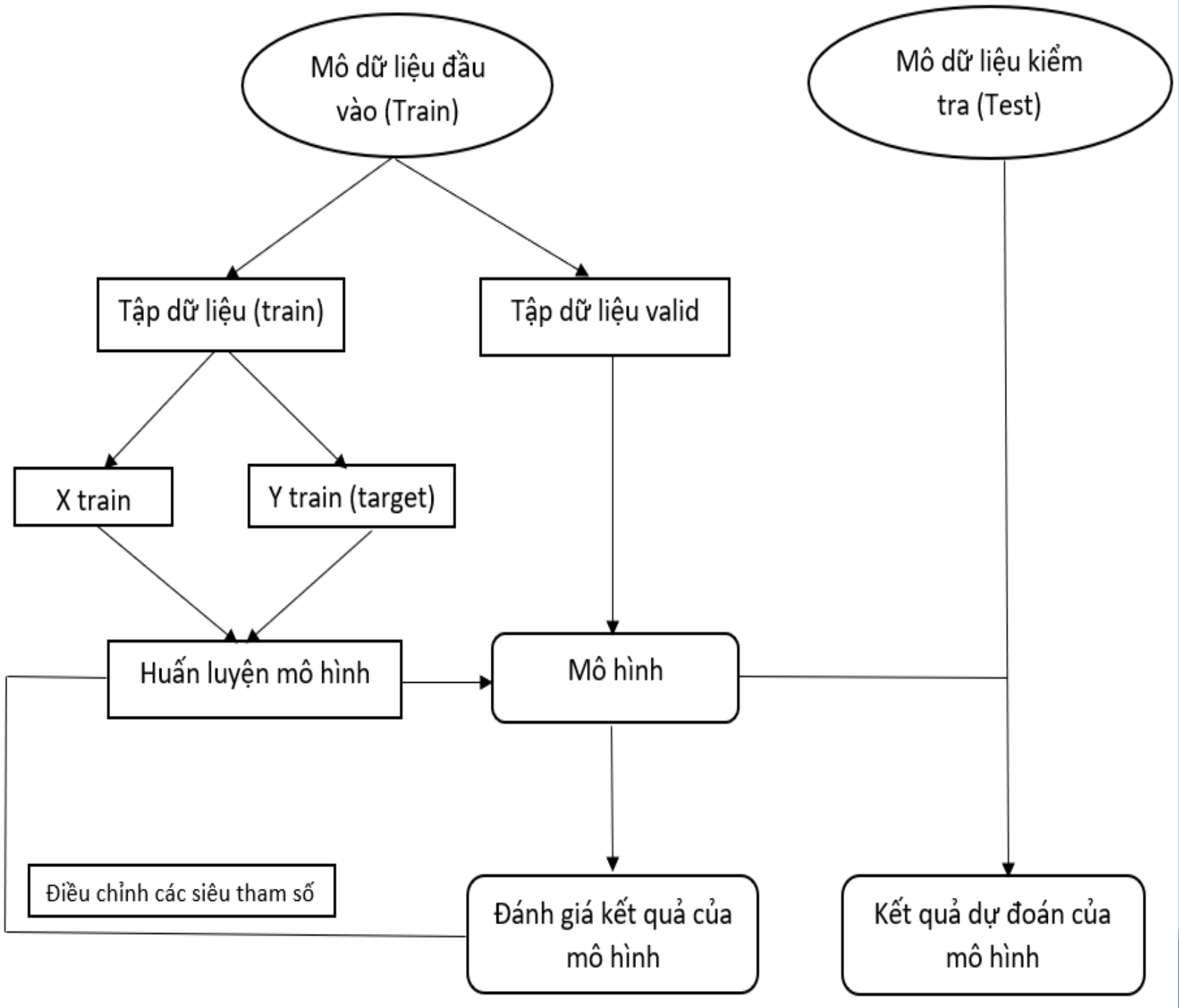
Nguồn ảnh: HopDock: A probabilistic search algorithm for decoy sampling in protein-protein docking - Scientific Figure on ResearchGate

Tổng quan quy trình chương trình

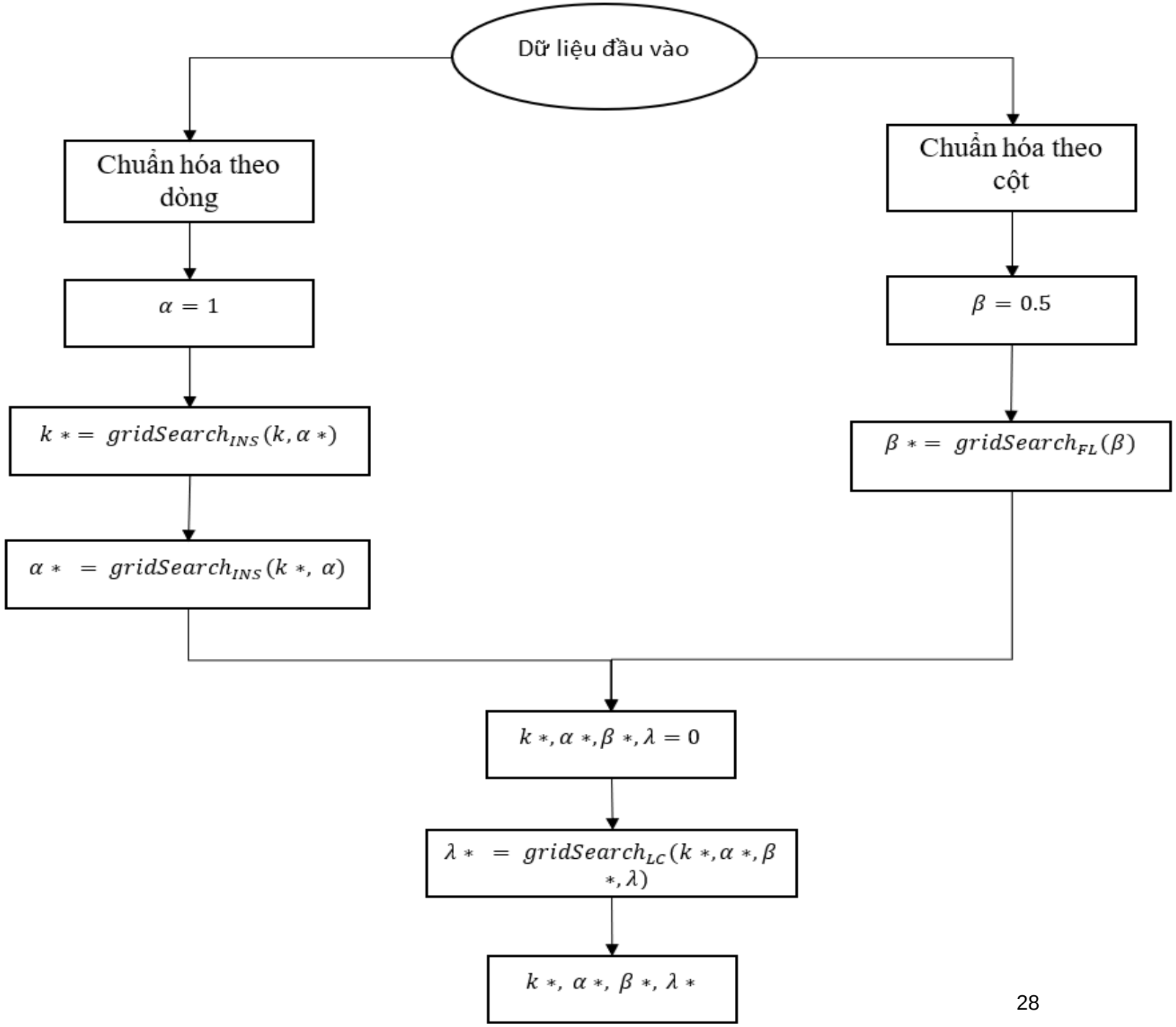


Thực nghiệm

Quy trình thực nghiệm



Quy trình huấn luyện mô hình



Thực nghiệm

Tìm kiếm lưới siêu tham số tốt nhất với tập dữ liệu và độ đo cho trước.

- Các tập có kích thước lớn: thực hiện kiểm tra chéo 10 phần.
- Các tập kích thước cực lớn: lấy mẫu 1000 hoặc 10000 phần tử đầu tiên trong tập kiểm tra.

Siêu tham số	Ý nghĩa	Miền giá trị
k	Số lân cận gần nhất cho dự đoán dựa trên đối tượng	{1,5,50,100,150,200,250,300,350}
α	Hệ số mũ cho độ tương đồng trên đối tượng	{0.5, 1, 1.5, 2}
β	Hệ số mũ biến đổi độ tương đồng trên đặc trưng	{0.5, 1, 1.5, 2}
λ	Trọng số kết hợp 2 thành phần dự đoán	$\{0, 0.1, 0.2, \dots, 0.9\}$

Mô tả dữ liệu

Bộ dữ liệu

	$\|X_{train}\|$	$\|X_{test}\|$	$\|F\|$	$\|y\|$	Lcard(T)	Mô tả
Large						
Medical	333	645	1449	45	1.25	Các tài liệu chứa văn bản lâm sàng miễn phí tại trung tâm y tế bệnh viện dành trẻ em
Corel5K	5000	500	499	374	3.53	Phân hiển thị việc phân loại ngữ cảnh như rừng, biển, bầu trời, …
Bibtex	4880	2515	1836	159	2.38	Hiển thị vấn đề về việc gán thẻ.
Delicous	12920	3185	500	983	19.04	Hiển thị vấn đề về việc gán thẻ.
IMDB-F	72551	48368	1001	28	1.97	Hiển thị thể loại phim dựa trên văn bản tóm tắt phim.
Extreme						
Wiki10	14127	6617	101890	30940	18.64	Bộ dữ liệu này hiển thị hơn 20.000 bài viết trên Wikipedia
Eurlex	15539	3809	5000	3956	5.3	Một bộ sưu tập các tài liệu về luật Châu Âu

Github: https://github.com/xfcdxg/mulan-lib

Website: http://manikvarma.org/downloads/XC/XMLRepository.html

Độ đo đánh giá

Các độ đo đánh giá

Độ chính xác dựa trên đối tượng (example based-accuracy) là độ đo thể hiện trung bình mỗi đối tượng kiểm tra có tỷ lệ dự đoán có nhãn i và dự đoán đúng là bao nhiêu.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|\mathbf{y_i} \cap \hat{\mathbf{y}_i}|}{|\mathbf{y_i} \cup \hat{\mathbf{y}_i}|}$$

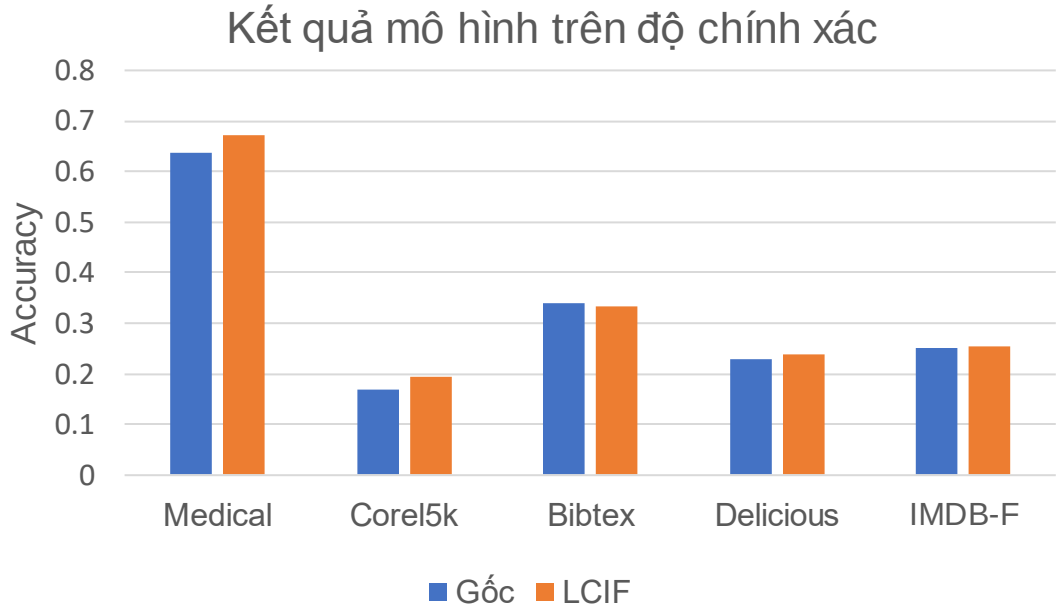
Hamming loss là phần của các nhãn sai trên tổng số nhãn.

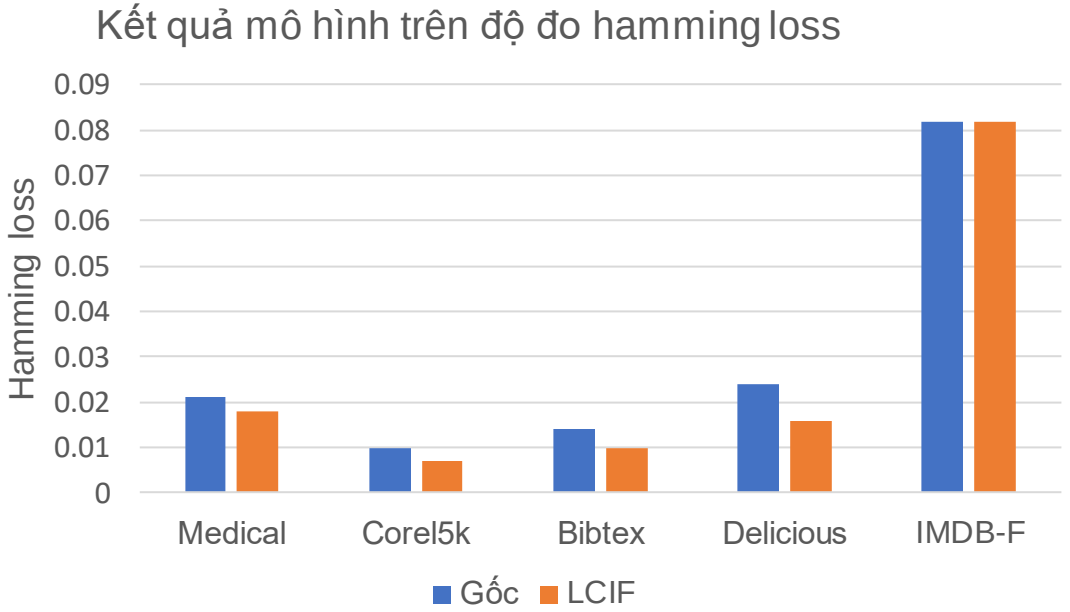
$$\text{Hamming loss} = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L}[I(y_j^{(i)} \neq \hat{y}_j^{(i)})]$$

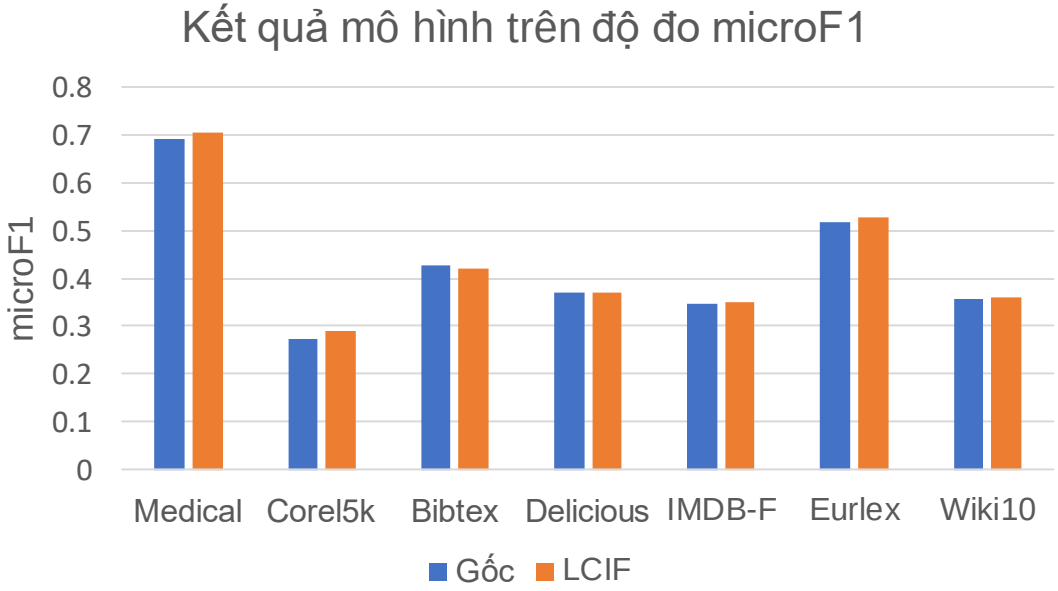
F1-score là trung bình điều hòa sự cân bằng giữa precision và recall.

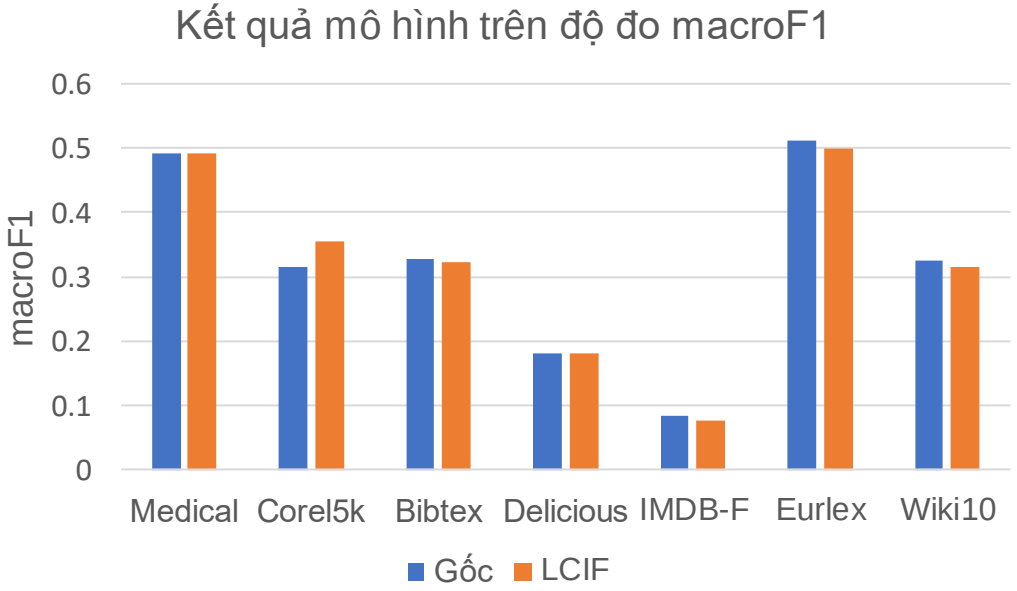
$$\text{Precision} = 2 \text{ x } \frac{\text{recall x precision}}{\text{recall+ precision}}$$

Kết quả


Kết quả mô hình trên độ chính xác
Accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Medical
Corel5k
Bibtex
Delicious
IMDB-F
Gốc
LCIF


Kết quả mô hình trên độ đo hamming loss
Hamming loss
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Medical
Corel5k
Bibtex
Delicious
IMDB-F
Gốc
LCIF


Kết quả mô hình trên độ đo microF1
microF1
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Medical
Corel5k
Bibtex
Delicious
IMDB-F
Eurlex
Wiki10
Gốc
LCIF


Kết quả mô hình trên độ đo macroF1
macroF1
0.6
0.5
0.4
0.3
0.2
0.1
0
Medical
Corel5k
Bibtex
Delicious
IMDB-F
Eurlex
Wiki10
Gốc
LCIF

Kết luận và hướng phát triển

Kết luận

Những gì đã đạt được trong đề tài nghiên cứu:

- Khảo sát các phương pháp đang được sử dụng hiện nay trong bài toán phân lớp đa nhãn.
- Đề xuất giải pháp mới: sử dụng phương pháp kết hợp dựa trên các đối tượng và đặc trưng đã cải tiến kết quả so với nghiên cứu trước đó.
- Áp dụng một số kỹ thuật mới như lấy ngưỡng threshold tự động, tối ưu hóa tham số, sử dụng độ tương đồng mới để hỗ trợ xây dựng mô hình hiệu quả hơn.

Hướng phát triển

Hướng phát triển trong tương lai:

▶Thực hiện tối ưu hóa thiết kế, cài đặt mô hình có hiệu năng cao hơn.

▶ Thử nghiệm các phương pháp tiền xử lý dữ liệu khác.

▶ Áp dụng những tiến bộ đạt được trong các thuật toán lọc cộng tác, có thể là một thuật toán tính độ tương đồng tiên tiến.

Cảm ơn quý thầy cô đã chú ý theo dõi

Phụ lục

Độ tương đồng cộng hưởng

- Độ nhất quán:

$$C(u, v, k_1) = \left(\sqrt{0.5 + 0.5\cos(\varphi_u - \varphi_v)}\right)^{k_1}$$

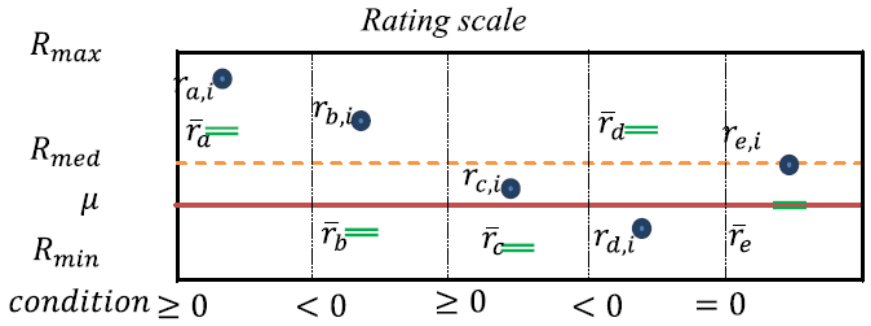
$$\varphi_u = \begin{cases} \frac{\pi}{R_{max} - R_{min}} \cdot \varphi_u^-, \ condition < 0 \\ \frac{\pi}{R_{max} - R_{min}} \cdot \varphi_u^+, \ condition \geq 0 \end{cases}$$

$$condition = \varphi_u^{base} \times (\bar{r}_u - \mu)$$

$$\varphi_u^+ = \frac{1}{1 + |\bar{r}_u - \mu|} \cdot \varphi_u^{base}; \ \varphi_u^- = 1 + \frac{|\bar{r}_u - \mu|}{R_{med}} \cdot \varphi_u^{base}$$

$$R_{med} = \frac{R_{max} - R_{min}}{2}$$

$$\varphi_u^{base} = (r_{u,i} - R_{med})$$



Ví dụ các trường hợp của $condition$

Nguồn ảnh: TAN, Zhenhua; HE, Liangliang. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle.

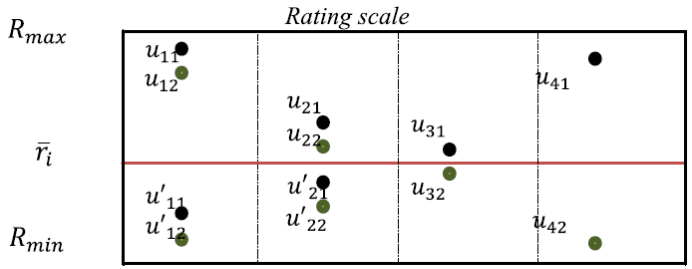
Độ tương đồng cộng hưởng

- Khoảng cách:

$$D(u,v,k_2,k_3) = \begin{cases} D^-(u,v,k_3), & (r_{u,i}-\bar{r}_i)(r_{v,i}-\bar{r}_i) < 0 \\ D^+(u,v,k_2), & (r_{u,i}-\bar{r}_i)(r_{v,i}-\bar{r}_i) \geq 0 \end{cases}$$

$$D^+(u,v,k_2) = \left(e^{-|r_{u,i}-r_{v,i}|} \times e^{\frac{1}{2}\times(|r_{u,i}-\bar{r}_i|+|r_{v,i}-\bar{r}_i|)}\right)^{k_2}$$

$$D^-(u,v,k_3) = (e^{-|r_{u,i}-r_{v,i}|})^{k_3}$$



- Hệ số Jaccard:

$$J(u,v,k_4) = \frac{|I_{u,v}|}{|I_u| + |I_v| - |I_{u,v}|}$$

Nguồn ảnh: TAN, Zhenhua; HE, Liangliang. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle.

Thực nghiệm

Ngưỡng điểm dự đoán (threshold): loại bỏ một số dự đoán $\hat{y}_{q,j} \in \hat{Y}$ thấp hơn ngưỡng.

$$\hat{Y}_t = \{h(x_q) = \{y_j \mid \hat{y}_{q,j} > t\}, \forall j \in L\}, \forall x_q \in X_q$$

Giá trị t được xác định sao cho:

$$argmin_t \left| lcard(\hat{Y}_t) - lcard(\hat{Y}) \right|$$

Tính trung binh số nhãn/đối tượng

$$lcard(D) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{L} \delta(d_{i,j})$$

Các độ đo đanh giá

- Example –base Accuracy.
- Hamming Loss.
- Micro-F1.
- Macro-F1.

Các độ đo đánh giá

MicroF1 là F1-score mà các giá trị TP, FP và FN được tính tổng trên tất cả giá trị thành phần tương ứng của các nhãn trong tập dữ liệu

$$F1_{micro} = \frac{\sum_{j=1}^{L} tp_j}{\sum_{j=1}^{L} tp_j + \frac{1}{2}\left(\sum_{j=1}^{L} fp_j + \sum_{j=1}^{L} fn_j\right)}$$

MacroF1 chỉ lấy trung bình không trọng số của các giá trị F1 tại mỗi nhãn.

$$F1_{macro} = \frac{1}{L} \times \sum_{j=1}^{L} \frac{tp_j}{tp_j + \frac{1}{2}\left(fp_j + fn_j\right)}$$

I. Kết quả trên bộ dữ liệu cơ sở

- KẾT QUẢ TỐT NHẤT TỪNG MÔ HÌNH TRÊN TỪNG ĐỘ ĐO.

	INS.KNN – RES	INS.KNN – cosine	LCIF - RES	LCIF - cosine
Accuracy				
Medical	0.585	0.563	**0.671**	0.636
Corel5k	0.187	0.163	**0.195**	0.17
Bibtex	0.341	**0.347**	0.332	0.341
Delicious	0.23	0.23	**0.24**	0.23
IMDB-F	0.25	0.247	**0.253**	0.25
Hamming Loss				
Medical	0.02	0.025	**0.018**	0.021
Corel5k	0.012	0.014	**0.007**	0.01
Bibtex	0.02	0.017	**0.01**	0.014
Delicious	0.025	0.025	**0.016**	0.024
IMDB-F	0.09	0.094	**0.082**	**0.082**
Micro F1				
Medical	0.656	0.622	**0.705**	0.69
Corel5k	**0.293**	0.266	0.291	0.274
Bibtex	0.417	0.426	0.42	**0.427**
Delicious	0.368	0.368	**0.371**	0.369
IMDB-F	0.35	0.341	**0.351**	0.346
Macro F1				
Medical	0.349	0.339	0.407	**0.492**
Corel5k	0.315	0.315	**0.354**	0.315
Bibtex	0.32	0.321	**0.328**	**0.328**
Delicious	0.181	0.18	**0.181**	**0.181**
IMDB-F	0.085	**0.09**	0.08	0.084