Collaborative Filtering Algorithms

Group 10

Yiyi Zhang; Hanying Ji; Jiaqi Dong; Mingming Liu

- Data Set
- Memory-Based Model
- Model-Based Model
- Comparison
- Conclusion

Part 1

Data Set

Data Set

Microsoft Web Data

The data records the use of www.microsoft.com by 38000 anonymous, randomly-selected users.

- implicit
- 1 for visit
- 0 for no visited
- Use Ranked Score to evaluate

$$R_a = \sum_j \frac{\max(v_{a,j} - d, 0)}{2^{(j-1)/(\alpha-1)}} \qquad R = 100\frac{\sum_a R_a}{\sum_a R_a^{max}}$$

EachMovie

61265 users entered a total of 2811983 numeric ratings on 1623 movies, i.e. about 2.4% entries are rated by zero-to-five stars.

- explicit
- Score from 1 to 6
- NA for no rated
- Use MAE to evaluate

$$S_a = \frac{1}{m_a} \sum_{j \in P_a} |p_{a,j} - v_{a,j}|$$

Assignment

Algorithm	Component	Variants	Data
Memory-based Algorithm	Similarity Weight	Pearson	1,2
		Spearman	1,2
		SimRank	1
	Variance Weighting	No	1,2
		Yes	1,2
	Selecting Neighbors	Weight Threshold	1,2
Model-based Algorithm	Cluster Models		2

Part 2

Memory-Based Algorithm

Memory-Based Algorithm

$$p_{a,i} = \overline{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \overline{r}_u) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}}$$

$$p_{a,i} = \overline{r}_a + \sigma_a * \frac{\sum_{u=1}^{n} \frac{r_{u,i} - \overline{r}_u}{\sigma_u} * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}}$$

Similarity Weighting

Pearson Correlation

$$w(a,i) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}}$$

Spearman Correlation

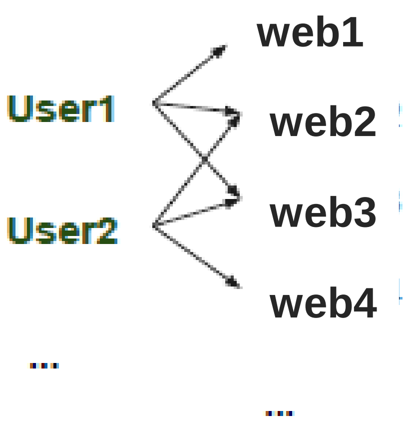
$$w_{a,u} = \frac{\sum_{i=1}^{m} (rank_{a,i} - \overline{rank_a}) * (rank_{u,i} - \overline{rank_u})}{\sigma_a * \sigma_u}$$

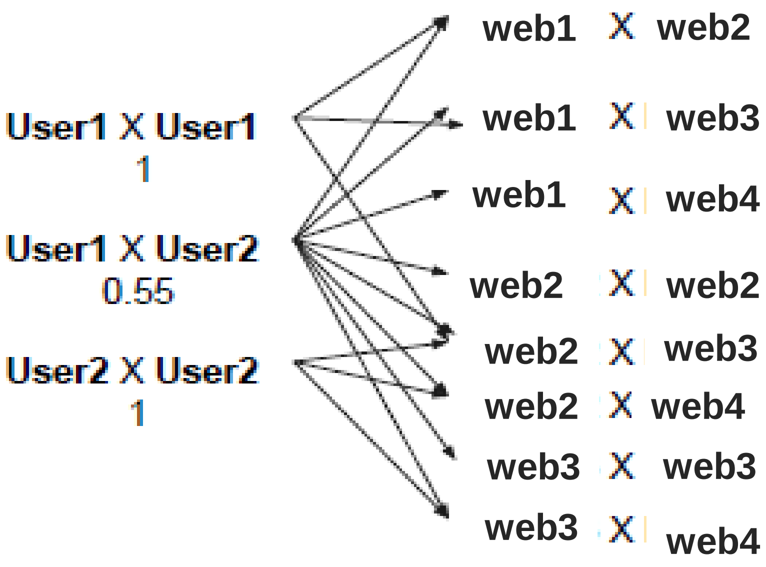
Similarity Weighting

Simrank

$$s_1(a,b) = \frac{C_1}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} s_2(O_i(a), O_j(b))$$

$$s_2(a,b) = \frac{C_2}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s_1(I_i(a), I_j(b))$$





Variance Weighting

Vairance Weighting
(Pearson Correlation)

$$w_{a,u} = \frac{\sum_{i=1}^{m} v_i * z_{a,i} * z_{u,i}}{\sum_{i=1}^{m} v_i}$$

Selecting neighbors

Selecting neighbors
(Weight Threshold)

$$W_{a,u} = \begin{cases} \omega_{a,u}, & \omega_{a,u} > threshold, \\ 0 & o.w. \end{cases}$$

Result – Data1

Rank Score		Z-Score				Deviation for Mean			
		Variance=OFF		Variance=ON		Variance=OFF		Variance=ON	
		Pearson	Spearman	Pearson	SimRank	Pearson	Spearman	Pearson	SimRank
					40.77057				40.77628
Weight Threshold	0.2	39.364175 (52.42%)	39.364175 (52.42%)	39.977511 (50.15%)		39.459138 (52.42%)	39.459138 (52.42%)	39.959964 (50.15%)	
	0.3	38.077085 (29.03%)	38.077087 (29.02%)	39.145602 (34.91%)		38.171786 (29.03%)	38.171798 (29.02%)	39.233054 (34.91%)	
	0.4	5.820478 (15.14%)	5.820477 (15.14%)	38.091526 (21.31%)		5.857705 (15.14%)	5.857706 (15.14%)	38.232782 (21.31%)	
	0.5	1.835299 (5.76%)	1.837463 (5.77%)	5.856577 (11.02%)		1.8487 (5.76%)	1.846338 (5.77%)	5.871136 (11.02%)	

Result – Data2

MAE		Z-Score			Deviation for Mean		
		Variance=OFF		Variance=ON	Variance=OFF		Variance=ON
		Pearson	Spearman	Pearson	Pearson	Spearman	Pearson
Weight Threshold	0	1.428475 (100%)	1.593824 (100%)	1.88334 (100%)	1.336934 (100%)	1.319426 (100%)	1.69374 (100%)
	0.2	1.137397 (57.83%)	1.137854 (57.81%)	1.145734 (50.10%)	1.135716 (57.83%)	1.136039 (57.81%)	1.139967 (50.18%)
	0.3	1.147375 (40.65%)	1.149654 (39.92%)	1.159057 (34.82%)	1.140594 (40.64%)	1.142161 (39.92%)	1.140594 (40.64%)
	0.4	1.16645 (24.28%)	1.171506 (23.15%)	1.182244 (21.21%)	1.149662 (24.28%)	1.1532 (23.15%)	1.149662 (24.28%)
	0.5	1.202854 (11.80%)	1.210711 (10.82%)	1.216701 (10.93%)	1.169631 (11.80%)	1.174899 (10.82%)	1.169631 (11.81%)

Part 3

Model-Based Algorithm

Model-Based Algorithm

Score Estimation

$$\mathbb{E}[V_b^{(i)}|v_j^{(i)}, j \in I(i)] = \sum_{k=1}^{5} k \cdot P\big(V_b^{(i)} = k|v_j^{(i)}, j \in I(i)\big),$$

$$= \frac{\sum_{c=1}^{C} P(\Delta_i = c) \cdot P(V_b^{(i)} = k|\Delta_i = c) \cdot \prod_{j\in I(i)} P(V_j^{(i)} = v_j^{(i)}|\Delta_i = c)}{\sum_{c=1}^{C} P(\Delta_i = c) \cdot \prod_{j\in I(i)} P(V_j^{(i)} = v_j^{(i)}|\Delta_i = c)},$$

Log-likelihood Function

$$l(\mu, \gamma|data) = \sum_{i=1}^{N} l_i(\mu, \gamma|data)$$

$$= \sum_{i=1}^{N} \log\Big[\sum_{c=1}^{C} \mu_c \cdot \prod_{j\in I(i)} P(V_j^{(i)} = v_j^{(i)}|\Delta_i = c)\Big],$$

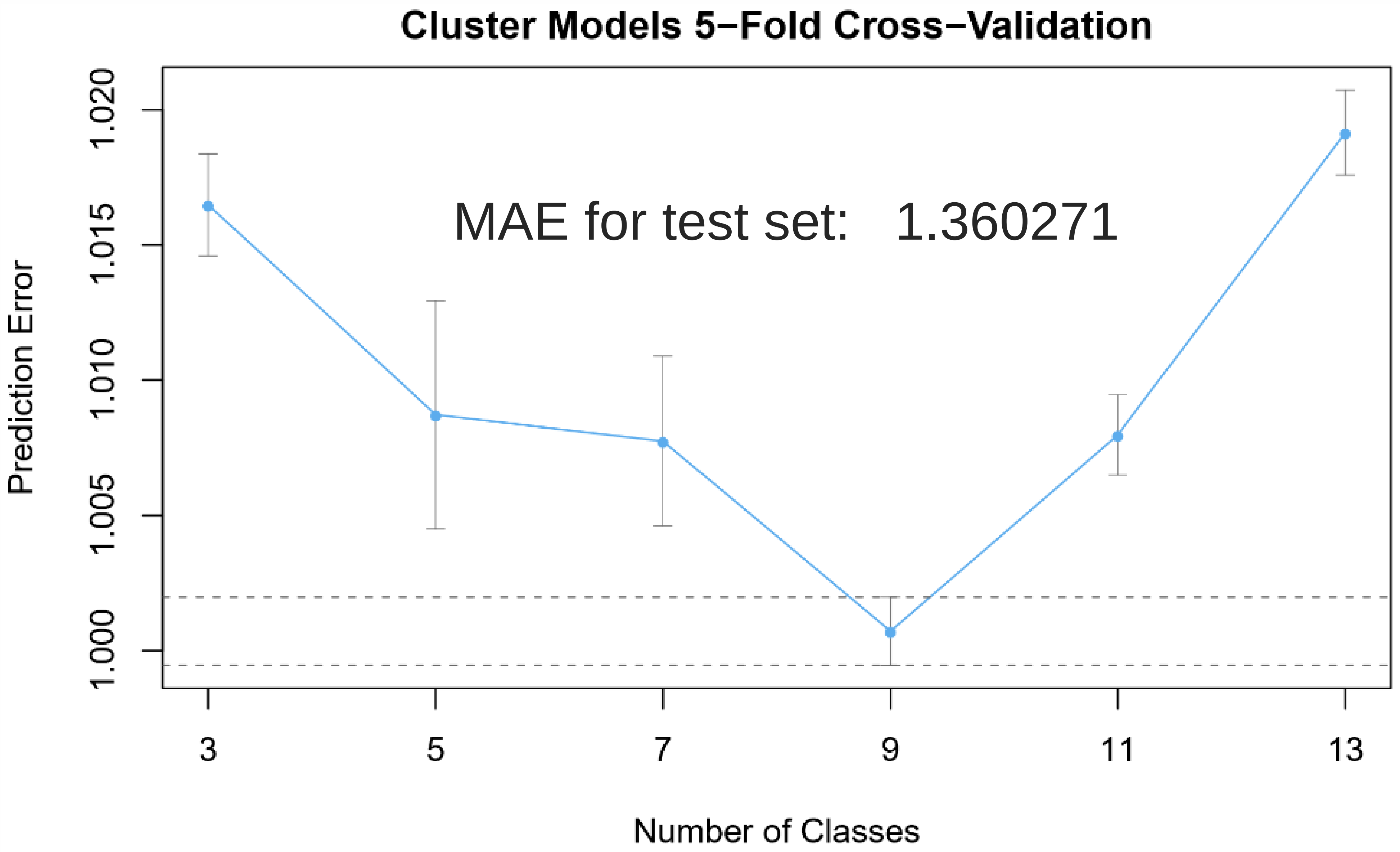
EM Algorithm

$$\hat{\pi}_i^c = \frac{\hat{\mu}_c \cdot \hat{\phi}_c(D(i))}{\sum_{c=1}^{C} \hat{\mu}_c \cdot \hat{\phi}_c(D(i))}$$

$$\hat{\mu}_c = \frac{\sum_{i=1}^{N} \hat{\pi}_i^c}{N}, \quad \text{for} \quad c = 1, \ldots, C$$

$$\hat{\gamma}_{c,j}^{(k)} = \frac{\sum_{i:j\in I(i)} \hat{\pi}_i^c \cdot \mathbb{I}(v_j^{(i)} = k)}{\sum_{i:j\in I(i)} \hat{\pi}_i^c}, \quad \text{for} \quad \forall c, j, k$$

Cross-Validation



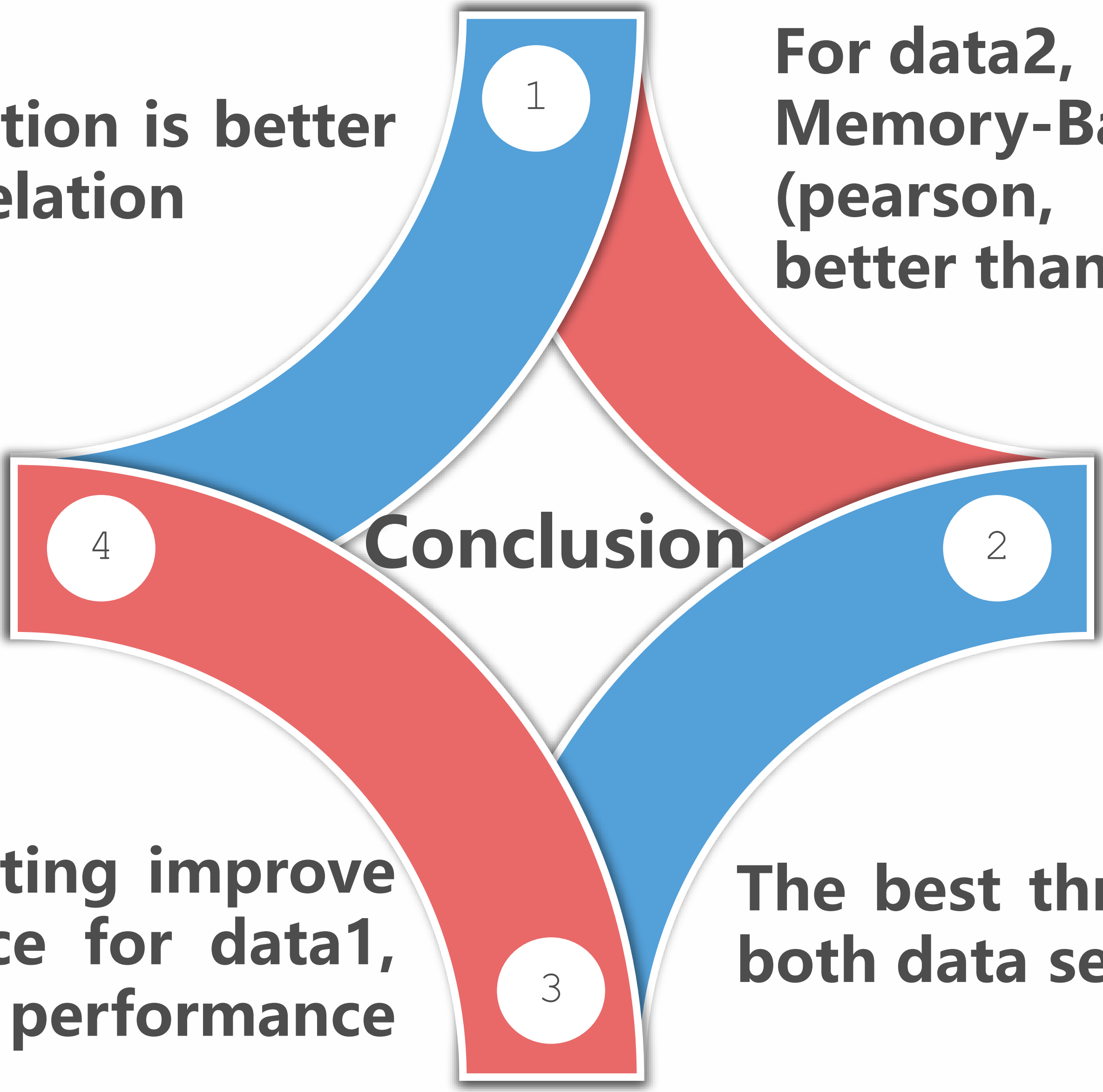
Comparison —— Data 2

	MAE
Cluster Model	1.360271
Memory-Based Model (Corrleation = Pearson Variance = No Threshold =0.2)	1.135716

Conclusion

For data1,
Simrank correlation is better than other correlation

For data2,
Memory-Based Model (pearson, threshold=0.2) is better than Cluster Model



The best threshold is 0.2 for both data sets

Variance Weighting improve the performance for data1, while lower the performance for data2

Thank You!

Group 10

Yiyi Zhang; Hanying Ji; Jiaqi Dong; Mingming Liu