The Minority Matters: A Diversity-Promoting Collaborative Metric Learning Algorithm

Shilong Bao[1,2], Qianqian Xu[4], Zhiyong Yang[1,2], Yuan He[6], Xiaochun Cao[7], Qingming Huang[2,3,4,5*]

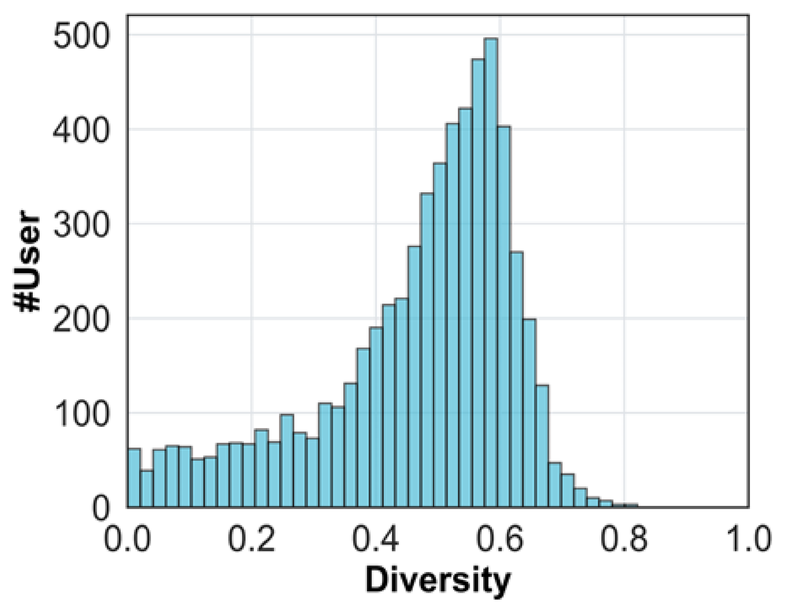
[1]Institute of Information Engineering, CAS [2]University of Chinese Academy of Sciences [3]BDKM, CAS

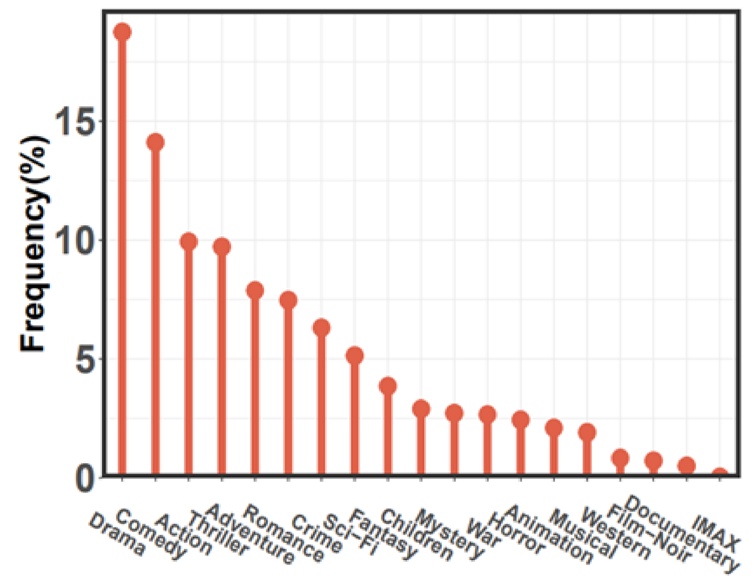
[4]Institute of Computing Technology, CAS [5]Peng Cheng Laboratory [6]Alibaba Group [7]Sun Yat-sen University

Motivation

Collaborative Metric Learning (CML) has recently emerged as a popular method in RS. However,

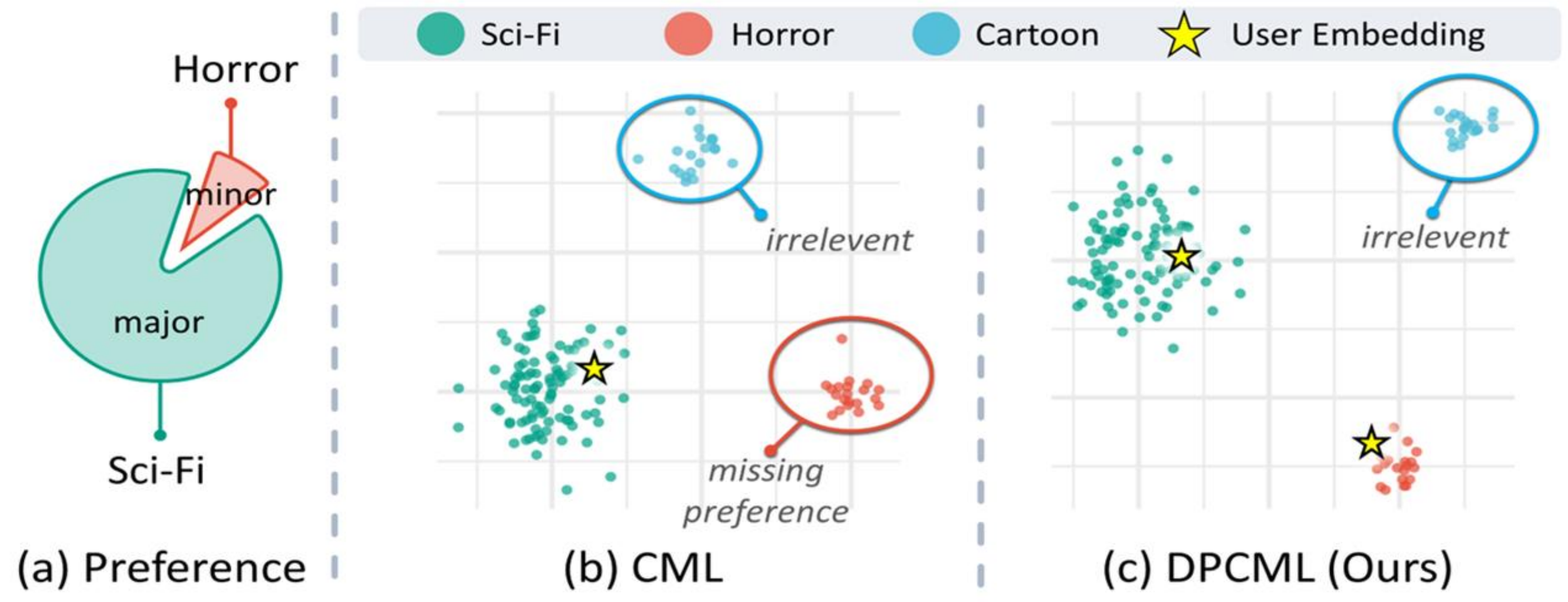
- the practical recommendation system mainly faces two-fold challenges, as shown in the following figure:
  a) Users often have multiple interests and preferences
  b) The item category distribution is highly imbalanced





An example visualization on MovieLens-10M dataset.

- Unfortunately, CML-based methods may induce preference bias and overlook the minority interest of users due to the limited expressive ability:



This work proposes an effective diversity-promoting algorithm to tackle the above challenges.

Diversity-Promoting Collaborative Metric Learning

- Each user u_i is represented by C different vectors, i.e.,

$$\boldsymbol{g}_{u_i}^c = \boldsymbol{P}_c^\top \boldsymbol{u}_i, \ \ \forall c, u_i, \ c \in [C], \ \ u_i \in \mathcal{U}$$

- Each item v_j is represented by one single vector, i.e.,

$$\boldsymbol{g}_{v_j} = \boldsymbol{Q}^\top \boldsymbol{v}_j, \ \ \forall v_j \in \mathcal{I}$$

- The user preference toward an item is defined as the minimum user-item Euclidean distance:

$$s(u_i, v_j) = \min_{c \in [C]} \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j}\|^2, \forall v_j \in \mathcal{I}$$

- Minimization goal for preference consistency

$$\hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-)$$

Safe Margin

$$\ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-) = \max(0, \lambda + s(u_i, v_j^+) - s(u_i, v_k^-))$$

Diversity Control Regularization Scheme

- Measure the representation diversity within the embedding sets of a given user, i.e.,

$$\delta_{\boldsymbol{g},u_i} = \frac{1}{2C(C-1)} \sum_{c_1, c_2 \in [C]} \|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2}\|^2$$

- Based on the fact that user's interests should not be too different, we put forward the following risk

$$\hat{\Omega}_{\mathcal{D},\boldsymbol{g}} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \psi_{\boldsymbol{g}}(u_i)$$

Minimum diversity

Maximum diversity

$$\psi_{\boldsymbol{g}}(u_i) = \max(0, \delta_1 - \delta_{\boldsymbol{g},u_i}) + \max(0, \delta_{\boldsymbol{g},u_i} - \delta_2)$$

Controlling proper diversity is crucial to good generalization!

Theoretical analysis

$$\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) = \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} + \eta \cdot \hat{\Omega}_{\mathcal{D},\boldsymbol{g}}$$

Theorem 1. Generalization Upper Bound of CML

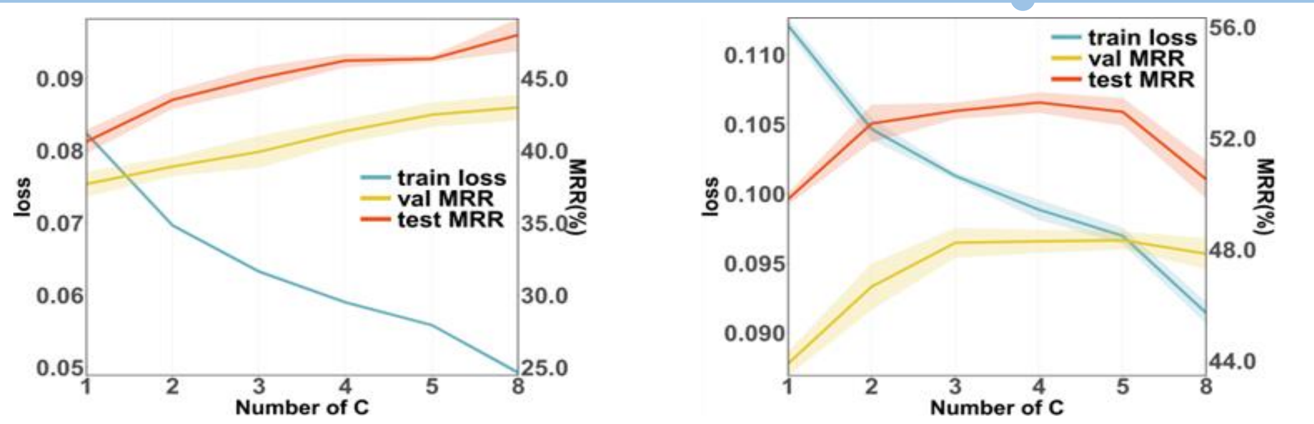
Let $\mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]$ be the population risk of $\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})$. Then, $\forall \ \boldsymbol{g} \in \mathcal{H}_R$, with high probability, the following inequation holds:

Population Risk

Empirical Risk

$$\mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \leq \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) + \sqrt{\frac{2d\log\left(3r\tilde{N}\right)}{\tilde{N}}}$$

It depends on the hypothesis space and data (not related to C)



DPCML could achieve a smaller empirical risk (related to vector number C)

DPCML enjoys a smaller generalization error than CML!

Experiments

	Type	Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
MovieLens-1m	Item-based	itemKNN	12.24	2.90	12.41	12.43	4.29	12.79	8.34	26.16
	MF-based	GMF	14.10	2.81	14.33	14.28	4.08	14.73	8.29	29.51
		MLP	13.95	2.78	14.22	14.06	3.98	14.56	8.30	29.39
		NeuMF	16.43	3.20	16.87	16.73	4.68	17.40	9.69	33.23
		M2F	8.61	1.84	9.36	7.60	2.30	8.67	2.95	20.40
		MGMF	17.38	3.51	18.08	17.63	5.05	18.52	10.12	35.15
	CML-based	UniS	17.56	3.71	17.89	18.34	5.60	18.79	12.40	35.77
		PopS	12.96	3.11	13.30	12.82	4.41	13.40	7.59	28.61
		2stS	21.07	4.84	21.35	21.81	7.07	22.29	14.42	40.36
		HarS	<u>24.88</u>	<u>5.86</u>	<u>25.38</u>	<u>24.89</u>	<u>8.25</u>	<u>25.77</u>	<u>15.74</u>	<u>45.15</u>
		TransCF	10.03	1.84	10.31	10.90	3.09	11.20	7.07	23.66
		LRML	17.15	3.52	17.56	17.45	5.12	18.08	10.42	34.36
		AdaCML	19.06	4.12	19.31	19.74	6.23	20.20	13.30	37.36
		HLR	21.10	4.80	21.53	21.61	7.06	22.28	13.95	40.71
	Ours	DPCML1	19.12	4.14	19.34	19.90	6.27	20.29	13.24	37.55
		DPCML2	**25.18**	**6.06**	**25.64**	**25.35**	**8.51**	**26.16**	**16.09**	**45.32**

Personal homepage: https://github.com/statusrank (with QR code)

Email: baoshilong@iie.ac.cn

Selected as oral presentation paper